2-3-2020

Deep Learning Systems for Automated Lesion Detection, Segmentation, and Classification in Mammography

Dina Abdelhafiz
University of Connecticut - Storrs, dina.abdelhafiz@uconn.edu

Deep Learning Systems for Automated Lesion Detection, Segmentation, and Classification in Mammography

Dina Abdelhafiz Hussein Abdelhafiz, Ph.D.

University of Connecticut, 2020

Breast cancer is the second leading cause of cancer deaths among women in the USA. Mammography is the preferred screening tool for breast cancer and accounts for the greatest contribution to the early detection of breast cancer. The detection of breast masses in mammogram (MG) images using deep learning (DL) systems is a challenging task due to the varying sizes, shapes, and textures of masses.

In this thesis, we propose a novel DL network called residual attention UNet (RAU-Net), the network pays attention to small lesions, and shows superior performance compared to the other state-of-the-arts DL models in detecting and segmenting masses, especially for heterogeneously dense and dense MG images. The proposed RAU-Net model achieves a mean dice coefficient index of 0.98 and mean intersection over union of 0.94. We propose a DL residual network for classification of MG images into benign and malignant that achieved accuracy of 0.95, and AUC of 0.98. We also propose a one-shot multi-input Siamese network that learns features from previous and current year MG images of the same patient to give a better assessment for current year MG images. The detection of mass tumors in dense tissues and, more generally, in dense breasts is

often considered more challenging due to the similar visual aspects of normal and abnormal dense tissues.

In this thesis, we present a training algorithm that we used to train various kinds of U-Net networks such as RCNN-UNet, AU-Net, RAU-Net,and UNet++ to generate density attention masks that automatically pays attention and gives more weight to tumors in dense regions of MG images. To train and test our models, we collected and pre-processed MG images that come with different resolutions from public repositories and MG images from UCONN health center.

In conclusion, we proposed DL systems for lesion detection, segmentation, and classification in mammography that can aid radiologists and serve as a second eye for them.

Deep Learning Systems for Automated Lesion Detection, Segmentation, and Classification in Mammography

Dina Abdelhafiz Hussein Abdelhafiz

B.S., Alexandria University, 2006

M.S., Alexandria University, 2011

M.S., University of Connecticut, 2019

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2020

APPROVAL PAGE

Doctor of Philosophy Dissertation

Deep Learning Systems for Automated Lesion Detection, Segmentation, and Classification in Mammography

Presented by

Dina Abdelhafiz Hussein Abdelhafiz, M.S.

Co-Major Advisor

Sheida Nabavi

Co-Major Advisor

Reda Ammar

Associate Advisor

Sanguthevar Rajasekaran

Associate Advisor

Clifford Yang

University of Connecticut

2020

DEDICATION

This thesis is dedicated to my loved Mum and Dad, who supported and encouraged me throughout my life.

To my sister Rania and my brothers Mohammed and Tarek, and my nephews Marwan and Moataz.

Mum, I wish you are proud of me at heaven. I will never forget your calls every-other hour to check on me. This Ph.D. long journey may have taken me far from you, but for a good reason.

Dad, you are a true leader, thanks for motivating me and keep me going all the time. Thanks for your technical and academic support all these years.

To all breast cancer fighters, "While you may not have chosen to have breast cancer, you still can choose to fight back and not to let breast cancer take away the motivation and the passion to achieve your dreams, this research is dedicated for you!" – Dina Abdelhafiz.

ACKNOWLEDGEMENTS

I take this opportunity to express my gratitude to my committee for their continued endless support and encouragement: Prof. Sheida Nabavi, Prof. Reda Ammar, Prof. Sanguthevar Rajasekaran, Prof. Cliff Yang and Prof. Jinbo Bi.

Prof. Sheida Nabavi, my co-major advisor, I am grateful for your guidance, encouragement, support and all long days revising my work and giving me scientific feedback. You were a big support for me as women to lead and do progress in science.

I cannot express enough thanks to my co-major advisor, Prof. Reda Ammar, for his support, professional advice, academic guidance and giving me many learning opportunities. Thank you!.

My thanks to my associated advisor, Prof. Sanguthevar Rajasekaran, for all the wonderful extensive lectures in algorithms and big-data. It's been a pleasure attending your classes.

Special thanks to my associated advisor, Prof. Cliff Yang, for guiding me through all my medical-related questions with answering me with details.

Prof Jinbo Bi, thank you for making the machine learning lectures awesome. I learned a lot!.

Mr. John, David, and Dan, the staff at the HPC center, I appreciate your technical support.

Rebecca, Joy, and Lara, the staff at the CSE department. Thanks for your friendly technical help.

My lab mates at NabaviLab, thanks for your feedback, arguments, and suggestions during our weekly meetings: Fatima Zare, Tianyu Wang, Abdelrahman Hosny, Thomas Davis, and Jun Bai.

Jim and Elaine Haggerty, my American family, thank you for all your support, care and love.

Finally, special thanks to Marwan Youssef, for all the happy five years. Thank you!.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

.

ABBREVIATIONS

The next list describes several abbreviations that will be later used within the body of this thesis.

ACC accuracy

ACR American college of radiology

AD architectural distortion

AI artificial intelligence

ANN artificial neural network

AUC area under the receiver operating curve

BACs breast arterial calcifications

BCDR breast cancer digital repository

BI-RADS breast imaging reporting and data system

BN batch normalization

Breast MRI breast magnetic resonance imaging

CAD computer-aided detection

CBIS-DDSM curated breast imaging subset of digital database for screening mammography dataset

CBMIR content based medical image retrieval

CC craniocaudal

CLAHE contrast limited adaptive histogram equalization

CNNs convolutional neural networks

CRF conditional random field

DA dense areas

DBT digital breast tomosynthesis

DDMS digital database for screening mammography

DICOM digital Imaging and Communications in medicine

DL deep learning

DM digital mammogram

DW discrete wavelet

E2E end-to-end

FCL fully connected layer

FCN fully convolutional network

FFDM fully field digital mammography

FN false negative

FNR false negative rate

FP false positive

FPR false positive rate

FROC free-response ROC

FSMs film screen mammograms

GLCM gray level co-occurrence matrix

GTM ground truth map

HGD histogram of the gradient divergence

HOG histogram of oriented gradients

ILSVRC imageNet large scale visual recognition competition

IRMA image retrieval in medical applications

KNN k-nearest neighbor

LDA linear discriminant analysis

MCs micro-calcifications

MG mammogram

MIAS mammographic image analysis society

ML machine learning

MLO mediolateral-oblique

MLP multilayer perceptron

NCCN the national comprehensive cancer network

NDA non-dense areas

PCA principal component analysis

PM pectoral muscle

PMD probability map of breast density

PNG portable network graphics

Pool pooling layer

R-CNN region-based network

ReLU rectified liner unit

ResNet residual network

ROC receiver operating characteristic curve

ROIs region of interests

SAE stacked autoencoder

SRCNN super resolution image reconstruction

SSL semi-supervised learning

SVM support vector machine

TL transfer learning

TN true negative

TNR true negative rate

TP true positive

TPR true positive rate

WHO world health organization

Chapter 1

Introduction

1.1 Motivations

Besides skin cancer, breast cancer is the most common cancer diagnosed among USA women. In 2017, breast cancer accounts for 30% of all new cancer diagnoses in women in the USA [200]. As of January 2018, there are more than 3.1 million women with a history of breast cancer in the USA. About 40,920 of them are expected to die in 2018, though death rates have been decreasing since 1989. The decrease in the mortality rates are thought to be the result of treatment advances, earlier detection through screening, increased awareness and new advances in early detection using computer-aided models (CAD) [204, 209]. Screen-film mammography (SFM) and full-field digital mammography (FFDM) are the primary imaging screening tools for the early detection of breast cancer. Radiologists visually search mammograms (MGs) for specific abnormalities. Mammography has contributed significantly to the reduction of the mortality rate through early detection of cancer [138]. However, the complexity of MGs and the high volume of exams per radiologist can result in false diagnosis [235]. Moreover, due to the low contrast, poor quality, and noisy nature of SFM MG images, a significant number of abnormalities are missed or misdiagnosed.

CAD models, which employ image processing techniques and pattern recognition theory, has

been introduced to provide an objective view to radiologists [138]. Studies have shown the effectiveness of CAD models; however, accurate detection of breast cancer has remained challenging [138]. In the last decade, traditional CAD models aim to aid radiologists in their interpretation of MGs, but in general, these techniques are outdated nowadays and cannot be applied to detect all types of breast abnormalities.

The extraction of breast mass in MG images using CAD systems is a challenging task due to the varying sizes, shapes, and textures of masses. The detection of masses in dense tissues and, more generally, in dense breasts is often considered more challenging due to the similar visual aspects of normal and abnormal dense tissues, which complicates the interpretation of mammographic images. Most of the traditional CAD systems need to have hand-crafted image features to be implemented. In our recent survey on traditional CAD models [9], it shows that traditional models have limitations in classifying MG images. Most of the traditional models depend on a prerequisite set of local hand-crafted features that cannot be generalized to work on a new data-set. Traditional CAD models consider limited feature types (e.g. texture features, shape features, and grey level intensity features), which require expert knowledge for selecting them. Poor feature extraction and selection cause challenges to build a successful classifier. These traditional systems may significantly reduce the amount of effort needed for the assessment of a lesion in clinical practice, while they may introduce a number of false-positives (FPs) that lead to unnecessary and discomforting biopsies. Recent studies show that CAD models cannot improve significantly the diagnostic accuracy of mammography [133]. The biggest challenge in using CAD for abnormality detection in MGs is the high false-positive rates (FPR). FPs result in patient anxiety, additional radiation exposure, unnecessary biopsies, high callback rates, increased health care costs, and additional

assessment [235]. In the USA, millions of women undergo screening mammography each year, as a result, even a small reduction in the FPR result in a widespread benefit [198]. The limitations of current CAD models indicate the need for new, more precise detection methods.

Recent advances in computational technologies, significant progress in machine learning and image processing techniques, and prevalence of digital MGs have opened up an opportunity to address the challenging issue of early detection of breast cancer using deep learning (DL) methods [4, 9]. Recently, DL methods, especially convolutional neural networks (CNNs) have gained lots of attention to CAD for MGs as they help overcome CAD systems' limitations [138, 252]. CNNs achieve higher detection accuracy than CAD models and help radiologists provide more accurate diagnosis by delivering quantitative analysis of suspicious lesions [9, 122, 169]. A recent research study shows that using DL methods drop the human error rate for breast cancer diagnoses by 85% [230]. Current CNN models are designed to improve radiologists' ability to find even the smallest breast cancers at their earliest stages alerting the radiologist to the need for further analysis [122, 230].

Recent studies used CNNs to generate a standard description of lesions, which can help radiologists in making a more accurate decision [122, 169]. Moreover, advances in CNNs can not only aid radiologists but also eventually make diagnosis systems to read MGs independently in the near future [122]. The development done in deep CNN to conduct different applications in mammography will save many lives. In the last few years, CNNs have led to breakthroughs in a variety of pattern recognition and classification problems for natural images due to the availability of big data repositories, fast graphical processing units, and the power of parallel and distributed computing [127, 129]. CNNs have also achieved state-of-the-arts performance for image classification,

lesion detection for mammography, and medical applications in general. Contrary to conventional CAD methods, which require a hand-crafted feature extraction stage that relies on domain knowledge, deep CNN methods adaptively learn the appropriate feature extraction process from the input data-sets with respect to the target output. CNNs extract local and global features from MG images, where the first layers of the network capture basic coarse features such as oriented edges, corners, textures, and lines while subsequent layers construct complex structures or global features. This eliminates the tedious process of investigating the discrimination ability of the features while facilitating the reproducibility of the methods.

Training a deep CNN model with a limited number of medical data is very challenging, which has been addressed by using transfer learning (TL) and augmentation (Aug.) techniques [127, 129, 234]. Studies show that CNN methods that compare images from left and right breasts [50] and also the craniocaudal (CC) and mediolateral oblique (MLO) view of each breast can improve the accuracy of detection and reduce the FPs, as discussed in [9]. CNNs have also been used in the risk assessment applications to increase the accuracy of early detection of breast cancer by radiologists.

In this thesis, We developed automated DL systems to precisely localize, segment, and classify mass lesions in MGs and address the above challenges. To develop a DL model with a competitive detection accuracy, First, we collected and pre-processed MG images that come with different resolutions from public repositories and UCONN health center (UCHCDM). Secondly, we started conducting a detailed review of the strengths, limitations, and performance of the most recent CNNs applications in analyzing MG images. Also, we studied the effect of popular pre-processing techniques on breast cancer classification and detection using state-of-the-art CNNs. We have developed a novel residual attention DL system for automated mass segmentation, detection, and

classification in MG images. The overall system precisely detects segments of mass lesions in MG images and classify the detected binary segmented lesions into benign or malignant. The developed DL system shows superior performance compared to the other state-of-the-art DL models, such as FCN, Dilated-Net, U-Net, SegNet, Faster R-CNN, and YOLO, in detecting and segmenting masses, especially for heterogeneously dense and dense MG images.

We focused on improving the developed DL CNN model using attention breast density technique which takes an MG image as input and generates a mask that estimates the breast density at the pixel level. The generated attention mask helps to identify and exploit the effective dense segments of MG images to support the proposed system in its detection and classification decision. We compared the performance of different state-of-the-arts models after and before using the proposed training technique. The models used in the previous density analysis are nested U-Net, recurrent residual U-Net, attention U-Net, and residual attention U-Net. Also, we propose a Siamese CNN model that incorporates prior-year MG images with current-year MG images. This can help to reduce false detections. The performance in terms of accuracy, precision, AUC, and F1, of the proposed model, is compared with a base-line network trained to classify images in current year images.

1.2 Outline of the Dissertation

This dissertation is organized as follows: In Chapter 2, we introduce the mammography images 2.1, structure of the breast 2.2, BI-RADS levels 2.3, abnormalities in mammogram 2.4, limitations of mammography 2.5, and breast cancer repositories 2.6. In Chapter 3, we give introduction about deep convolutional neural networks (DCNNs) architecture 3.1, popular CNNs 3.2, and pop-

ular performance metrics for validations of the CNN models 3.3. We summarize and discuss the research impact of the proposed approaches as follows:

(i) ***Illustrate the advances, challenges and applications of deep convolutional neural networks for mammography*** (Chapters 4, 5, and 6). The main goal of this aim is to contact a detailed review of the strengths, limitations, and performance of the most recent CNNs applications in analyzing MG images. In this survey, we summarizes 83 research studies (Table 6.2) for applying CNNs on various tasks in mammography [4, 8, 9] (Chapters 5 and 6). This survey focuses on finding the best practices used in these research studies to improve diagnosis accuracy. Also, it provides a deep insight into the architecture of CNNs used for various tasks. Furthermore, it describes the most common publicly available MG repositories and highlights their main features and strengths.

The mammography research community can utilize this survey as a basis for their current and future studies. Specifically, CNNs are used in mammography for lesion localization and detection, risk assessment, image retrieval, and classification tasks. CNNs also help radiologists providing more accurate diagnosis by delivering precise quantitative analysis of suspicious lesions. The given comparison in this survey among common publicly available MG repositories guides the community to select the most appropriate database for their application(s), more details are discussed in section 2.6. Moreover, this survey lists the best practices that improve the performance of CNNs including the pre-processing of images and the use of multi-view images (Chapter 4). Also, other listed techniques like transfer learning (TL), data augmentation, batch normalization (BN), and dropout are appealing solutions to reduce overfitting and increase the generalization of the CNN models. Finally, this survey

identifies the research challenges and directions that require further investigations by the community (Chapter 6) [4, 8, 9].

(ii) ***Discuss the effect of pre-processing on breast cancer detection using CNNs.*** The main goal of this aim is to study the effect of pre-processing on breast cancer detection using CNNs. We have carefully chosen several pre-processing filters for contrast enhancement and noise removal, which are popular in the medical field. We investigated the influence of these pre-processing filters on the classification of MG images using CNNs into three-classes: normal, benign, and malignant [5], more details are discussed in Chapter 7.

(iii) ***Developing CNN for automated mass segmentation in mammography.*** The main goal of this aim is to develop a deep learning model to precisely segment mass lesions in MGs and extracts low-level and high-level features from MGs [6], more details are discussed in Chapter 8.

(iv) ***Implementing residual attention deep learning system for automated mass segmentation and classification in mammography.*** The main goal of this aim is to propose a residual deep learning system for automatic mass segmentation and classification in mammography. The overall proposed system consists of two cascaded parts: 1) a residual attention U-Net model (RAttU-Net) to precisely segment mass lesions in MG images, followed by 2) a ResNet classifier to classify the detected binary segmented lesions into benign or malignant, more details are discussed in Chapter 9.

(v) ***Density attention-based convolution neural network for breast cancer detection in MGs.*** The detection of cancer in dense tissues and, more generally, in dense breasts is often considered

more challenging due to the similar visual aspects of normal and abnormal tissues, which complicates the interpretation of mammographic images. The density of MGs makes mass detection very challenging since masses can be hidden in dense MGs. For the above reasons, we argue that deep learning used for lesion detection should first evaluate the spatial distribution of the dense tissues. The main goal of this aim is to estimate the breast density at the pixel level while using only image-wise ground truth from the BI-RADS scale. Our goal is to generate an attention breast density mask, identifying pixels associated with the tissue that contributed to the density class. These generated attention masks to identify and exploit the effective dense segments of MGs images to support the CNN detection and classification decision. The approach is premises on the hypnosis that the there is benefit from amplifying the influence of dense regions, compared to treat a whole MG the same, especially for the heterogeneously dense and scattered regions. Attention masks of breast density will be extracted automatically using a residual attention U-Net model. To target the above challenges, the generated attention maps will help the main CNN used for detecting lesions to pay attention more to these regions, see results in Chapter 10.

(vi) ***Incorporation of previous years MGs in breast cancer detection.*** Radiologists compare MLO and CC views, look for asymmetry and evaluate changes with respect to prior year MGs. Findings that haven't changed from older MGs are likely to be benign tumors and are signs of FP detection. In contrast, abnormal cells grow at a more rapid rate and begin to invade surrounding tissues, they form cancerous tumors. These cancerous tumors have different shapes and sizes than in prior images of the same patients. However, the appearance of normal breast tissues doesn't change much across the years. The main goal of this aim is

to develop a deep learning model that is trained on both images and mimic radiologists in comparing prior and current MG images to improve the detection accuracy (Chapter 11).

(vii) ***Generative adversarial network for data augmentation of breast masses in Mammogram Images.*** We present our future work in Chapter 12.

1.3 Publications

1.3.1 Conferences

My thesis is based on the content of the following peer-reviewed conference and journal papers that are accepted and published with primary authorship which includes:

- **D. Abdelhafiz**, S. Nabavi, R. Ammar, and C. Yang, "Survey on deep convolutional neural networks in mammography," in Computational Advances in Bio and Medical Sciences (ICCABS), 2017 IEEE 7th International Conference on, pp. 1–1, IEEE, 2017.

- **D. Abdelhafiz**, S. Nabavi, R. Ammar, and C. Yang, "The effect of pre-processing on breast cancer detection using convolutional neural networks,"Poster session presented at the meeting of the IEEE International Symposium on Biomedical Imaging, Washington, DC, 2018.

- **D. Abdelhafiz**, S. Nabavi, R. Ammar, C. Yang and J. Bi, "Convolutional Neural Network for Automated Mass Segmentation in Mammography," 2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), Las Vegas, NV, 2018, pp. 1-1.

- **D. Abdelhafiz**, S. Nabavi, R. Ammar, and C. Yang, "Residual Deep Learning System for Mass Segmentation and Classification in Mammography." In 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB '19), September 7–10, 2019, Niagara Falls, NY, USA. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3307339.3342157

1.3.2 Journals

- **D. Abdelhafiz**, S. Nabavi, R. Ammar, C. Yang, "Deep convolutional neural networks for mammography: advances, challenges and applications," BMC Bioinformatics. 2019Jun;20(11):281.

- **D. Abdelhafiz**, S. Nabavi, R. Ammar, C. Yang, "Additional file 1 of Deep convolutional neural networks for mammography:advances, challenges and applications." BMC Bioinformatics.2019 Jun;20(11):281.

- **D. Abdelhafiz**, S. Nabavi, R. Ammar, C. Yang, J. Bi, "Convolutional Neural Network for Automated Mass Segmentation in Mammography," BMC Bioinformatics.

Chapter 2

Mammography and Mammograms Repositories

2.1 What is Mammography?

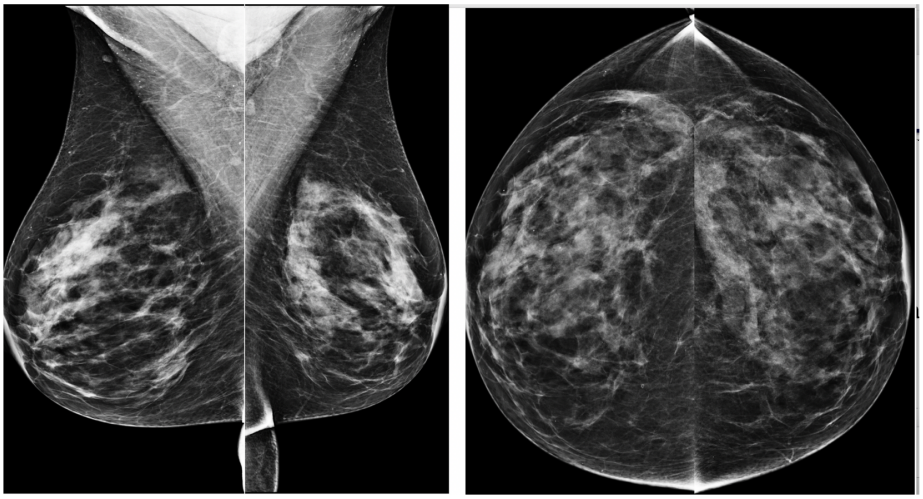
Breast cancer is the second most common cause of cancer deaths among women in the United States after lung cancer. As of January 2019, there are more than 3.1 million women with a history of breast cancer in the USA.

Screening mammography is used as the main imaging modality for the early detection of breast abnormalities. Studies have shown that it is the only method of breast imaging that consistently contributed to the decrease of breast cancer-related mortality. According to the American Cancer Society, the female breast cancer death rate declined by 40% from its maximum in 1989 to 2016 [67, 199] through early detection using mammography screening.

Mammography is an x-ray imaging method used to examine the breast for the early detection of cancer and other breast diseases. Imaging with x-rays involves exposing a part of the body to a small dose of ionizing radiation to produce pictures of the inside of the body. X-rays are the most frequently used form of medical imaging. It is used as both a diagnostic and screening tool. Screening mammography is the main imaging modality used for early detection of breast abnormalities. Studies have shown that it is the only method of breast imaging that consistently contributed to the

decrease of breast cancer-related mortality [170, 179]. According to the American Cancer Society, the female breast cancer death rate declined by 40% from its maximum in 1989 to 2016 through early detection using mammography screening [67, 199]. Mammogram (MG) images include the recording of two views for each breast (left and right breast): the craniocaudal (CC) view, which is a top to bottom view, and a mediolateral oblique (MLO) view, which is a side view [68, 205] (see Fig. 2.1). The Mammography images can be acquired on x-ray film, such as a film screen mammogram (FSM), or in digital formats, such as with digital mammography (full-field digital mammography (FFDM). Breast cancer can manifest on mammography in several ways: abnormal areas of lumps or tumors called masses, small white spots called calcifications (MCs), architectural distortion (AD), asymmetry, or any combination of these. Besides the primary aim of detecting breast cancers in screening mammograms, it is necessary to lower false-positive rates (FPR). False positives rates lead to patient anxiety, higher expenses, additional radiation exposure, unnecessary biopsies, high unnecessary callback rates [37].

Radiologists examine MG images to search for areas or types of tissue that look different from normal tissue. Radiologists look at the size, shape, and contrast of a mass, as well as the edges or margins, which can indicate the possibility of malignancy (i.e. cancer). They also look for tiny bits of calcium, called microcalcifications, which show up as very bright specks on a mammogram. When possible, the doctor reading a mammogram of a case will compare it to the patient's prior-year MG images. This helps radiologists find small changes that could be signs of cancer or know how the lesion developed within the breast. Radiologists find it very essential to have all the patients prior year exams, as it can show whether a mass or calcification has changed/non-changed for prior years. A non-change in the shape or size of the tumor would mean that it's likely not



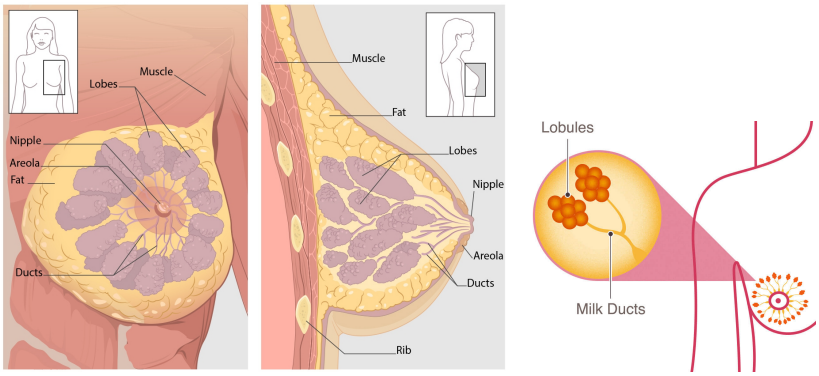
(a) Right MLO view, left MLO view (b) Right CC view, left CC view

Fig. 2.1: A craniocaudal (CC) view, which is a top to bottom view, and a mediolateral oblique (MLO) view, which is a side view.

cancer and a biopsy is not needed.

2.2 Structure of the Breast

Cancer occurs as a result of mutations, or abnormal changes, in the genes responsible for regulating the growth of cells and keeping them healthy. Breast cancer is a cancer that develops from breast tissue. Signs of breast cancer may include a lump in the breast, a change in breast shape, dimpling of the skin, fluid coming from the nipple, a newly-inverted nipple, or a red or scaly patch of skin. The breast composed from lobules (which are the glands that produce milk), ducts (the passages that drain milk from the lobules to the nipple), fatty and connective tissue, blood and lymphatic vessels, as in Figs. 2.2a, 2.2b and 2.2c. Breast cancers might begin in the ducts and called 'ductal cancers' while others start in the lobular and are named 'lobular cancers'. Less commonly, breast cancer can begin in the stromal tissues, which include the fatty and fibrous connective tissues of the breast. A tumor can be benign or malignant. Benign tumors are not considered cancerous: their cells are close to normal in appearance, they grow slowly, and they do not invade nearby tissues or spread to other parts of the body. Malignant tumors are cancerous. Malignant cells eventually can spread beyond the original tumor to other parts of the body and invade nearby healthy breast tissue. When a cancerous tumor settles in the mammary glands and the cancerous cells do not spread to other parts of the body, the breast cancer is considered 'non-invasive' or 'in situ'.



(a) Front view. (b) Side view. (c) Lobules.

Fig. 2.2: Structure of the breast. Photo credit(https://www.cdc.gov)

2.3 What is BI-RADS?

BI-RADS is the breast imaging reporting and data system classification that is proposed by the American College of Radioloogy (ACR) (http://www.acr.org). BI-RADS is widely used classification system for radiologist who report mammogram findings. The BI-RADS Atlas is widely used reference to improve communication between physicians, and provides standardized mammographic reporting, breast imaging terminology, a report organization and a classification system. It also describe the breast composition, any significant finding and its associated features using standardized terminology and conclude to a final assessment category. Based on level of malignancy, the lesions can be classified to one of six BI-RADS categories as described in Table 2.1.

Category 0 is defined as a diagnosis that needs to be combined with other imaging. Category 1 is defined as no lesions or negative findings. Category 2 is defined as benign lesion without suspicious characteristics. Category 3 is defined as benign possible with less than 2% malignant probability. Category 4 is defined as suspicious lesion with 2% to 95% malignant probability that is recommended for biopsy. Category 5 is defined as highly suspected of malignancy, with more than 95%

malignant probability. Category 6 is defined as known malignancy or pathologically proven to be malignant. Because of the wide range of malignancy probability, category 4 is divided into three subcategories: 4A, 4B and 4C, with 2:10%, 10:50% and 50:95% malignancy probability, respectively. Researchers in literature take the images in category 1 and label it as normal, and combine the images of categories 2:3 and label it as benign and combine categories 4:6 as malignant.

Table 2.1: BI-RADS assessment categories.

Category	Definitions
Category 0:	Exam is not conclusive or incomplete. Needs additional imaging evaluation and/or prior mammograms for comparison.
Category 1:	Negative / normal.
Category 2:	Benign findings.
Category 3:	Probably benign findings.
Category 4:	Suspicious abnormality.
Category 5:	Highly suspicious of malignancy.
Category 6:	Known biopsy with proven malignancy.

2.4 What are Kinds of Abnormality in Mammogram?

The most common findings seen on mammography are masses, calcifications (Calc), architectural distortion of breast tissue or a combination of them. Figure 2.3 shows MG images that contain a combination of mass and Calc lesions. The breast composition (breast density) which is not a cancer is used as a risk factor of breast cancer.

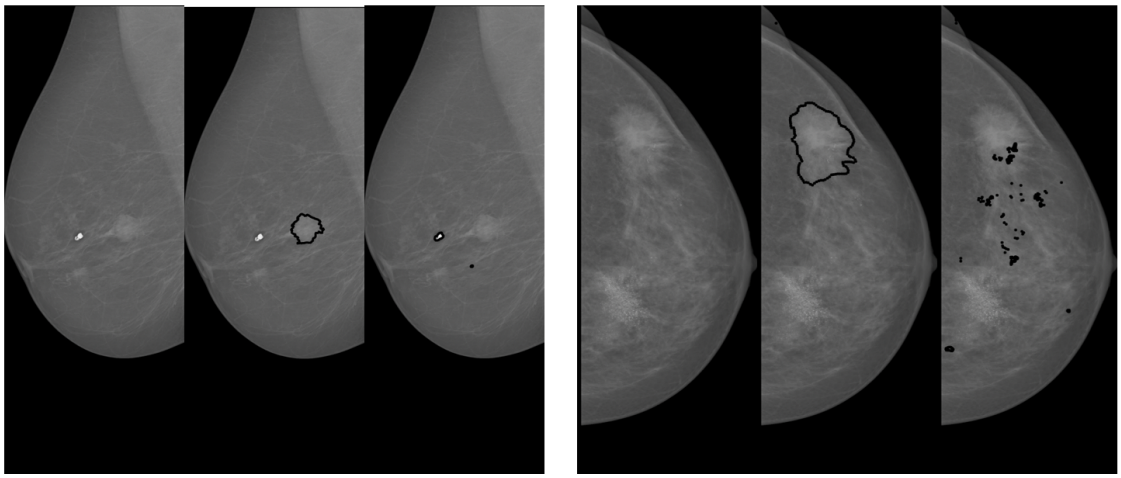


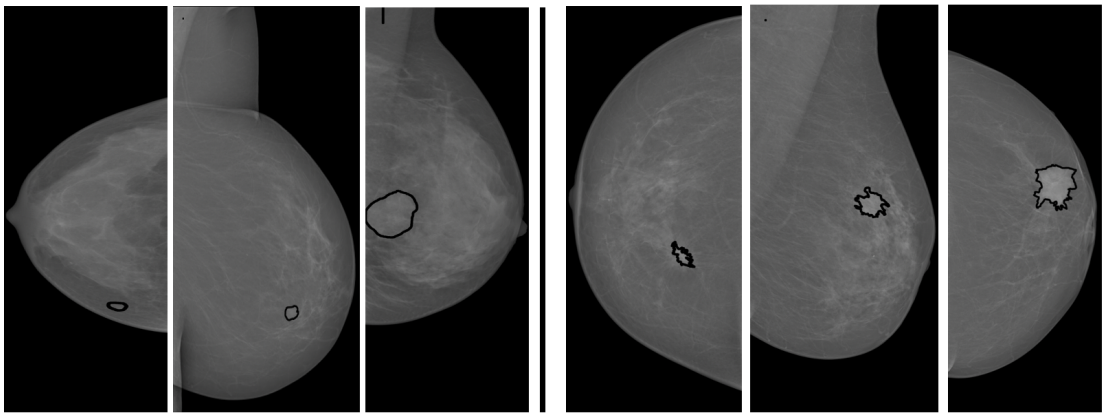
Fig. 2.3: (a) and (b) show two different breasts that have a combination of mass and Calc findings in the same breast.

2.4.1 Mass(es)

A mass is the same as a lump or a tumor, it is a space occupying 3D lesion seen in two different views, MLO and CC views. It can be a combined by calcifications. If a potential mass is seen in only a single projection, it is called a asymmetry until its three-dimensionality is confirmed. A Mass is described by its shape, margins and density as seen in Fig. 2.4 and described in Table 2.2. A circumscribed oval and round masses are usually benign, while Ill-defined, or spiculated shape suggests a greater likelihood of malignancy (see Fig. 2.4).

2.4.2 Calcifications

Calcifications are tiny mineral deposits within the breast tissue. They look like small white spots on a mammogram. They may or may not be caused by cancer. There are two types of calcifications, Macrocalcifications, and Microcalcifications. Macrocalcifications are coarse, bigger calcium



(a) Benign cases. (b) Malignant cases.

Fig. 2.4: (a) show benign cases, (b) show malignant cases.

Table 2.2: The terminology used to describe masses.

Shape:	Oval:	A mass that is elliptical or egg-shaped.
	Round:	A mass that is spherical, ball-shaped, circular or globular in shape.
	Lobular:	A mass that has contours with undulations.
	Irregular:	The lesion's shape cannot be characterized by any of the above.
Margins:	Circumscribed (Well-Defined):	The margins are sharply defined with an abrupt transition between the lesion and the surrounding tissue.
	Obscured:	One which is hidden by superimposed or adjacent normal tissue and cannot be assessed any further.
	Microlobulated:	The margins undulate with short cycles producing small undulations.
	Indistinct or Ill-defined:	The poor definition of the margins raises concern that there may be infiltration by the lesion.
	Spiculated:	The lesion is characterized by lines radiating from the margins of a mass.
Density	High, equal, low or fat-containing.	

deposits that are most likely due to changes caused by aging of the breast, old injuries, or inflammation. These deposits are related to non-cancerous conditions and don't need to be checked with a biopsy. On the other hand, Microcalcifications are specks of calcium in the breast. Microcalcifications seen on a mammogram takes more attention than Macrocalcifications, but they don't always mean that cancer is present. The shape and layout of Microcalcifications help the radiologist judge if it is malignant or benign. In the Atlas edition 2003, calcifications are classified by morphology and distribution either as benign, intermediate concern or high probability of malignancy. The distribution of calcifications is at least as important as morphology. The distribution includes diffuse, regional, grouped, linear, segmental grouping. These descriptors are arranged according to the risk of malignancy as in Table 2.3

Table 2.3: The terminology used to describe calcification.

Morphology of calcifications:	Benign:	The calcifications which includes skin, vascular, coarse or "popcorn", large rod like, round or punctate (< 1mm), rim, dystrophic, milk of calcium and suture.
	Suspicious morphology:	The calcifications include amorphous (small and/or hazy in appearance), coarse heterogeneous (between 0,5 mm and 1 mm), fine pleomorphic (usually < 0,5 mm), fine linear or fine-linear branching (usually < 0,5 mm)
Distribution of calcifications:	Diffuse:	Distributed randomly throughout the breast.
	Regional:	Occupying a large portion of breast tissue > 2 cm greatest dimension.
	Grouped/cluster:	Few calcifications occupying a small portion of breast tissue: lower limit 5 calcifications within 1 cm and upper limit a larger number of calcifications within 2 cm.
	Linear:	Arranged in a line, which suggests deposits in a duct.
	Segmental:	Suggests deposits in a duct or ducts and their branches.

2.4.3 Architectural Distortion

Architectural distortion (AD) are findings that represent unilateral deposits of fibroglandulair tissue not conforming to the definition of a mass. All types of asymmetry have different border contours than true masses and also lack the conspicuity of masses. Asymmetries appear similar to other dis-

crete areas of fibroglandulair tissue except that they are unilateral, with no mirror-image correlate in the opposite breast. The asymmetries description of Architectural distortion according to Atlas'latest edition includes asymmetry, focal asymmetry, global asymmetry, developing asymmetry, as in Table 2.4.

Table 2.4: The terminology used to describe the asymmetries of architectural distortion.

Asymmetry:	which as an area of fibroglandulair tissue visible on only one mammographic projection, mostly caused by the superimposition of normal breast tissue.
Focal asymmetry:	Visible on two projections, hence a real finding rather than superposition, this has to be differentiated from a mass.
Global asymmetry:	Consists of asymmetry over at least one-quarter of the breast and is usually a normal variant.
Developing asymmetry	New, larger and more conspicuous than on a previous examination.

2.4.4 Breast Composition

Breast density or breast composition is based on how fibrous and glandular tissues are distributed in your breast, versus how much of your breast is made up of fatty tissue. Breast density is a strong indicator of breast cancer, which shows the possibility of the detection of abnormalities in MGs. Dense breasts have a higher risk of developing breast cancer as breast tissue is made up of ducts and lobules, and the more ducts and the more lobules there are in the breast tissue, the denser the breast will appear or the whiter it will appear on mammography. Because of that density, sometimes cancers can be obscured or hidden. Cancers tend to be white on mammograms so if your breast tissue is dense and it is hard for radiology to find cancers on a dense mammogram. The breast density decreases with the woman's age, reflecting the gradual substitution of fibroglandular tissue by fat tissue. The American College of Radiology (ACR) introduced the Breast Imaging Reporting and Data System (BI-RADS) classification [9] (see Table 2.5). The four categories of this density system are: BI-RADS (A) indicates a fatty breast; BI-RADS (B) scattered fibroglandular densities;

BI-RADS (C) a heterogeneously dense breast and BI-RADS (D), the highest level, an extremely dense breast that could obscure a lesion (see Fig. 2.5).

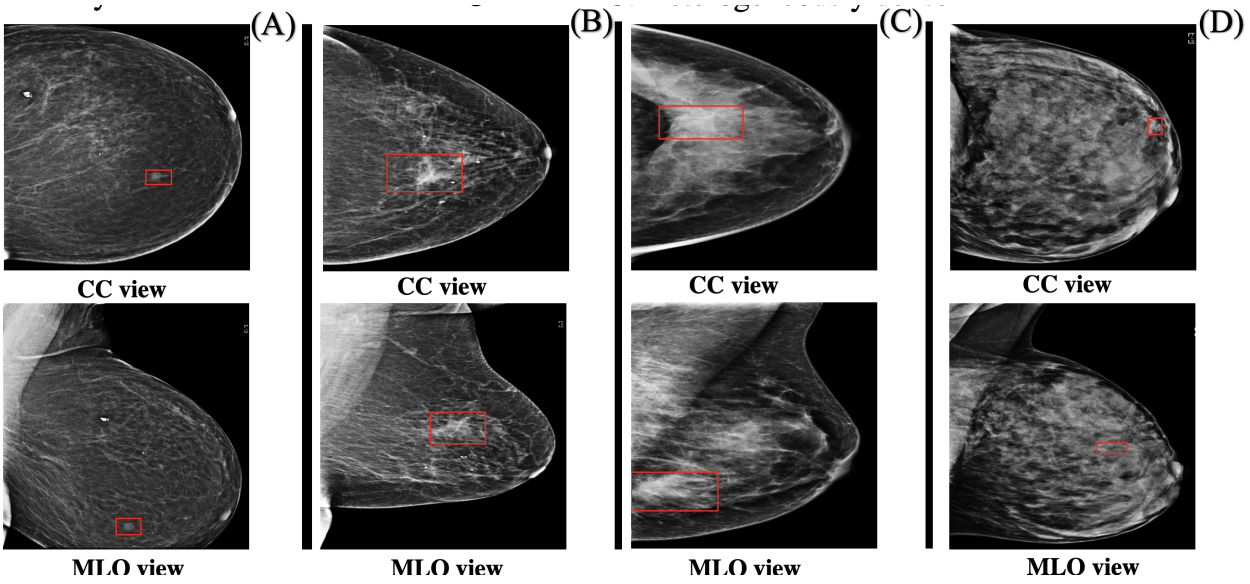


Fig. 2.5: The four categories of breast density are: BI-RADS (A) indicates a fatty breast; BI-RADS (B) scattered fibroglandular densities; BI-RADS (C) a heterogeneously dense breast and BI-RADS (D), the highest level, an extremely dense breast.

Table 2.5: The assignment of the breast composition.

BI-RADS class	Density (%)	Breast density	Remarks
A	0-25	Entirely fatty tissue	-
B	26-50	Scattered fibroglandular tissue	-
C	51-75	Heterogeneously dense tissue	Obscure small masses.
D	76-100	Extremely dense tissue	Lowers the sensitivity of mammography.

2.5 What are the Limits of Mammography for Radiologist and Medical Researchers?

For certain types of breasts, mammograms can be difficult to interpret by a radiologist or using CAD methods. This is because there is a wide variation in breast tissue density among women. Denser breasts are difficult to diagnose and make it harder for radiology to find cancers on a mammogram. The sensitivity of mammography in detecting cancer in dense breasts can vary over a wide range. Denser breasts is a risk factor for missed cancers and both false-positive and false-negative among young women who have dense breasts. For most of the CAD detection and classification methods used in literature to reduce false-positive results, it is highly recommended to do image enhancement for the mammogram images before applying any CAD methods.

2.6 Breast Cancer Digital Repositories

Mammographic databases play an important role in the development of algorithms aiming at the detection and diagnosis of mammography lesions and serve for testing and evaluation of these algorithms, also, as ground truth for the researchers. The amount of data needed to train a DL network is massive compared to the data needed to train traditional neural networks. The availability of comprehensive annotated databases is critical for advancing DL development in medical imaging.

2.6.1 Available Databases for Mammography

There are common publicly available databases for MGs: the Mammographic Image Analysis Society (MIAS) database [210], Digital Database for Screening Mammography (DDSM) [50], IN-

breast database [155], Breast Cancer Digital Repository (BCDR) [142], Image Retrieval in Medical Applications (IRMA) [160]. These databases present a wide variability of patients'cases and a mixture of normal, benign, and malignant cases. Annotations include the 'ground truth', the location and boundaries of the lesion with the outline marking performed by an imaging specialist. Some of these databases are publicly available and some are restricted to individual organizations.

Table 2.6 compares the publicly available MG databases according to the origin, the number of images, size of images, views (CC, MLO), digital or film database, the format of images, resolution of images, and the distribution of normal, benign and malignant images. Other databases used in literature are private and restricted to individual organizations [15, 17, 44, 70, 75, 103, 108, 115, 121, 154, 188, 212]. The public databases present a wide variability of patients' cases and a mixture of normal, benign, and malignant cases. Annotations include the location and boundaries of the lesions performed by imaging specialists. The public repositories have collected film screen MGs (FSMs) [142, 160, 210], and/or digital mammography (FFDM) [1, 142, 155, 160] with different resolutions. Digital MG images are usually saved in the Digital Imaging and Communications in Medicine (DICOM) format that gathers not only the image but also some related meta-data as in [1, 142, 155]; however, some databases use different formats [93, 142, 160, 210]. In the following subsection, we will talk about the famous databases recently used in literature and make a comparison between them in Table 2.6.

MIAS [210]: The Mammographic Image Analysis Society (MIAS) database is a publicly available database of mammographic data in the UK. despite being the smallest and one of the oldest databases, it is still widely used in literature till now. MIAS consists of 322 digitized film mammographic images (161 pairs), where 207 images represent normal, while 64 and 51 images referred

to as benign and malignant cases respectively. All of the images belong to the MLO view of the breast. MIAS also includes radiologist's "ground truth" markings on the locations of any abnormalities that may be present and types of suspicious regions, ACR breast density rating, and the severity of abnormality (BI-RADS). The database is originally digitized at 50-micron pixel edge and has been reduced to 200-micron pixel edge and padded/clipped so that all the images are 1024×1024. All images exist in a "Portable Gray Map" (PGM) format. The images of the MIAS database are of low resolution and have strong noise. MIAS includes four kinds of abnormalities (architectural distortions, stellate lesions, circumscribed mass, and calcifications). Although images are still is reachable at (http://peipa.essex.ac.uk/info/mias.html) , it is no longer supported. Despite all the previous drawbacks, it has been widely used in literature until now [3, 71, 104].

DDSM [50]: The Digital Database for Screening Mammography (DDSM), is a huge resource for publicly available breast cancer database mammographic images. DDSM is used in many studies [3, 32, 33, 40, 43, 55, 65, 112, 113, 119, 135, 143, 192, 195, 214, 218, 244, 258]. It is a collaborative effort between Massachusetts General Hospital, Sandia National Laboratories and the University of South Florida Computer Science and Engineering Department. DDSM is accessible at (http://http://marathon.csee.usf.edu/Mammography/Database.html). The database contains approximately 2,500 studies (Normal, Mass, Architectural distortion, Calcification). Each study includes two images of each breast (CC and MLO), along with some associated patient information (age at the time of the study, ACR breast density rating, subtlety rating for abnormalities, ACR keyword description of abnormalities) and image information (scanner, spatial resolution, ...). Images containing suspicious areas have associated pixel-level "ground truth"

information about the locations and types of suspicious regions marked made by experienced mammography. Moreover, the Region of Interest (ROI) annotations for the abnormalities in the DDSM images indicate a general position of lesions, without precise segmentation of them. The data-set is available as a compressed archive that requires decompression into a raw data format (RAW) and conversion to the original Portable Pixelmap format(PNM) based on the original image dimensions. The images are organized into 12 normal, 15 malignant and 16 benign volumes with a total of 8,752 images representing 2,620 cases. DDSM database has been used extensively in the literature to train deep learning models [4, 9].

INbreast Database [155]: where fully digitalized images were acquired at a breast center located in a university hospital (Centro Hospitalar de S. Joao [CHSJ], Breast Centre, Porto) with the permission of the Portuguese National Committee of Data Protection and Hospital's Ethics Committee. INbreast has a total of 115 cases (410 images) from which 90 cases are from women with both breasts affected (four images per case) and 25 cases are from mastectomy patients (two images per case). Several types of lesions (masses, calcifications, asymmetries, and distortions) were included along with some associated patient information. Accurate contours made by specialists are also provided in XML format. The strengths of the INbreast-relies on the fact that it was built with full-field digital mammograms, it presents a wide variability of cases and is made publicly available together with precise annotations. This database is gaining more attention nowadays in CAD methods for breast cancer detection and classification.

BCDR [142]: The Breast cancer digital repository (BCDR), is still in construction public repository. it is a collaborative effort between INEGI, FMUP-CHSJ- University of Porto, Portugal, and CETA-CIEMAT, Spain and Aveiro University. The BCDR database is subdivided into two differ-

ent repositories: 1- a film mammography-based Repository (BCDR-FM) and 2- a full-field digital mammography-based repository (BCDR-DM). BCDR is a compilation of Breast Cancer anonymous patients'cases annotated by expert radiologists containing clinical data (detected anomalies, breast density, BI-RADS classification, etc.), lesions outlines, and image-based features computed from CC and MLO mammography image views. The BCDR-FM is composed by 1,010 (998 female and 12 male) patients cases (with ages between 20 and 90 years old), including 1,125 studies, 3,703 MLO and CC mammography incidences and 1044 identified lesions clinically described (820 already identified in MLO and/or CC views). With this, 1,517 segmentations were manually made and BI-RADS classified by specialized radiologists. MLO and CC images are grey-level digitized mammograms with a resolution of 720 $\times$ 1,168 pixels and a bit depth of 8 bits per pixel, saved in the TIFF format. The BCDR-DM, still in construction, it is composed of 724 (723 female and 1 male) Portuguese patients' cases (with ages between 27 and 92 years old), including 1,042 studies, 3,612 MLO and/or CC mammography incidences and 452 lesions clinically described (already identified in MLO and CC views). With this, 818 segmentations were manually made and BI-RADS classified by specialized radiologists. The MLO and CC images are grey-level mammograms with a resolution of 3328 $\times$ 4084 or 2560 $\times$ 3328 pixels, depending on the compression plate used in the acquisition (according to the breast size of the patient). The bit depth is 14 bits per pixel and the images are saved in the TIFF format.

UCHCDM [253, 254]: The University of Connecticut Health Center (UCHCDM) is a private mammographic database belongs to the University of Connecticut. UCHCDM is still in the progress of data collection and more mammographic images will be available for CAD analyses soon. UCHCDM database contains 128 cases with a total of 870 mammograms (Normal, Mass,

Architectural distortion, Calcification). Both normal and malignant cases have successive scans from prior years and the current year. Each case study includes two images (CC and MLO) of each breast. All mammographic images are annotated in a descriptive text file with known pathology (Normal, Mass, Architectural distortion, Calcification) and circled at the locations of cancers (if any) on a separate key image. These annotations are the ground truths for model training and testing.

IRMA [160]: Image Retrieval in Medical Applications (IRMA) is developed from the union of the Mammographic Image Analysis Society Digital Mammogram Database (MIAS), The Digital Database for Screening Mammography (DDSM), the Lawrence Livermore National Laboratory (LLNL), and routine images from the Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen. These databases are of different resolutions and sizes. Using the IRMA code, standardized coding of tissue type, tumor staging, and lesion description was developed according to the American College of Radiology (ACR) tissue codes and the ACR breast imaging reporting and data system (BI-RADS). Disregarding the resolution, this resulted in a total of 10,509 reference images, and 6,767 images are associated with an IRMA contour information feature file. The database is freely available for research purposes. The ROI annotations for these databases are more precise making them more accurate for supervised DL methods.

CBIS-DDSM [132]: Curated Breast Imaging Subset of DDSM is an updated and standardized version of the Digital Database for Screening Mammography (DDSM). The CBIS-DDSM collection includes a subset of the DDSM data selected and curated by a trained mammographer. With a total of 4,067 images decompressed and converted to DICOM format. Updated ROI segmentation and bounding boxes, and pathologic diagnosis for training data are also included. The images are

read from DDSM and stored as 16-bit gray-scale DICOM files. Figure 2.6 shows a MG image from the MIAS, DDSM, BCDR, INbreast, UCHCDM data-sets.

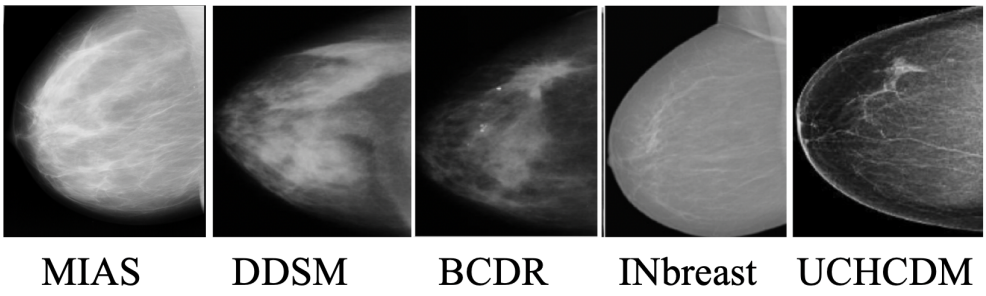


Fig. 2.6: Different mammogram databases.

2.6.2 Which Database is Suitable for your CAD Model?

This question can be answered if we analyzed the strengths and limitations of these databases are summarized in Table 2.7. Considering the number of publicity available image files, the DDSM database surpasses all the other available databases with quantity, and are excessively used in ANN and deep learning neural networks. But the drawback of this huge database is that DDSM images are saved in non-standard compression files that require the use of decompression code moreover, the ROI annotations for the abnormalities in the DDSM indicate the general position of lesions without precise segmentation for them. Low resolution and strong noise are drawbacks of MIAS images. The MIAS database lacks in quality, offering only 8-bit images, MIAS is an old database published in 1994 and no more supported. Another drawback of MIAS is that it contains a limited number of images (exactly 321) which make it of use with ANN and SVM but not deep learning algorithms. The INbreast database is gaining more attention nowadays in CAD methods for because of its high resolution and accurate segmentation of lesions, however, its size is still limited with limited variations of masses [25, 40, 58–60, 106, 236, 258]. The IRMA Project is a

combination of number of databases of different resolution and sizes, IRMA project corrected the counter information for these databases making it more accurate for CAD methods. With 10,509 images could be of great benefit to the deep learning methods. BCDR and UCHCDM are promising databases but still in their development phase. BCDR has been used in few studies [26,48,95,111]. CBIS-DDSM is an updated version of the DDSM database however it has a 1318 image of original 10480 images and is good for ANN methods.

2.6.3 Availability of Data-sets

The DDSM data-set is available online at (http://marathon.csee.usf.edu/Mammography/Database.html). The INbreast data-set can be requested online at (http://medicalresearch.inescporto.pt/breastresearch/index.php/Get_INbreast_Database). The breast cancer digital repository (BCDR) data-set can be requested online at (https://bcdr.eu). The MIAS database is available online at (http://peipa.essex.ac.uk/info/mias.html). The IRMA dataset can be requested online at (http://irma-project.org/index_en.php).

2.7 Summary

While mammography is the best screening tool for breast cancer available worldwide.Interpretations of mammograms can be challenging because a normal breast looks different for each woman. Because some breast cancers are hard to visualize, a radiologist may want to compare the image to views from previous examinations. Increased breast density has attracted attention in recent studies for multiple reasons, including increased breast density makes it difficult to see cancer on mammography. Increased breast density may increase the risk of getting breast cancer.

Table 2.6: Comparison between widely used databases in literature.

DB	# Cases	#Image	Image Size	Origin	Views	File Type	File Format	BI-RADS	Breast Density	Bits/Pixel	Patient Age	Mass Shape	Margins	GT	Lesion Type
DDSM	2620	10480	3118×5001	USA	MLO/CC	Film	LJPEG	Yes	Yes	12 bits	Yes	Yes	Yes	Chain code	All Kind
									ACR		(31-89)				
IRMA	-	10509	Several sizes	Germany	MLO/CC	Film	PNG	Yes	Yes	12 bits	Yes	Yes	Yes	Several	All Kind
						Digital			ACR		(30-90)				
INbreast	115	410	3328×4084	Portugal	MLO/CC	Digital	DICOM	Yes	Yes	16 bits	Yes	Yes	Yes	Central Radius	All Kind
			2560×3328								(20-85)				
MIAS	161	321	1024×1024	UK	MLO	Film	PGM	No	Yes	8 bits	No	Yes	Yes	Circle	All Kind
									(Not ACR)						
CBIS	-	1318	3118×5001	USA	MLO/CC	Digital	DICOM	Yes	Yes	12 bits	Yes	Yes	Yes	Chain code	All Kind
											(20-89)				
UCHC	141	1098	3328×4096	USA	MLO/CC	Digital	DICOM	Yes	No	12 bits	No	Yes	Yes	JPEG	All Kind
BCDR-F01	190	362	720×1168	Spain	MLO/CC	Film	TIF	Yes	Yes	8 bits	Yes	Yes	Yes	Chain code	All Kind
									ACR		(28-82)				
BCDR-F02	98	188	720×1168	Spain	MLO/CC	Film	TIF	Yes	Yes	8 bits	Yes	Yes	Yes	Chain code	All Kind
									ACR		(28-82)				
BCDR-F03	344	736	720×1168	Spain	MLO/CC	Film	TIF	Yes	Yes	8 bits	Yes	Yes	Yes	Chain code	All Kind
									ACR		(28-82)				
BCDR-D01	64	143	3328×4084	Spain	MLO/CC	Digital	DICOM	Yes	Yes	14 bits	Yes	Yes	Yes	Chain code	All Kind
			2569×3328						ACR		(23-88)				
BCDR-D02	50	200	3328×4084	Spain	MLO/CC	Digital	DICOM	Yes	Yes	14 bits	Yes	Yes	Yes	Chain code	All Kind
			2569v3328						ACR		(34-89)				
BCDR-DN01	50	456	3328×4084	Spain	MLO/CC	Digital	DICOM	Yes	Yes	14 bits	Yes	-	-	-	All Kind
			2569×3328						ACR		(31-89)				

Table 2.7: A summary for the strengths and limitations of the DDSM, IRMA, INbreast, MIAS, and BCDR databases.

Database	Strength	Limitation
DDSM	Big widely used database.	Non-standard format.
	Shape variations of different lesions.	Not precise position of lesions.
IRMA	Accurate position of lesions.	Non-standard format.
INbreast	High resolution.	Limited size.
	Accurate position of lesions.	Limited mass shape variations.
	Standard file format.	
MIAS	Still widely used.	Old database.
		No more supported.
		Limited size.
		Images are of low resolution.
		Has MLO view only.
BCDR	Different resolutions.	Limited size.
	Accurate position of lesions.	
	Standard file format.	
	Still in their development phase.	

Chapter 3

Deep Convolutional Neural Networks Architectures

3.1 Convolutional Neural Networks (CNNs)

In fact, DL is not a new idea, which even dates back to 1940s [129, 191] for medical images. Shallow layer CNNs were used to investigate breast cancer in 1995 [44, 232]. Famous CNNs such as Alex-Net [127], ZF-Net [248], GoogLeNet [215], VGG-Net [201] and ResNet [92] have brought about breakthroughs in processing images. Alex-Net architecture is extensively used in medical imaging for breast cancer detection. DL is a subset of machine learning that requires a huge number of labeled data to train the models. The term "deep" usually indicates the number of hidden layers in neural networks, e.g. ResNet has a depth of 152 layer which is 8× deeper than VGG-Net. Since 2012, CNNs have become more popular and have attracted more attention because of the increasing computing power, availability of lower-cost hardware, open-source algorithms, and the rise of big data [127].

3.1.1 Building Blocks of Convolutional Neural Networks

The structure of CNNs is very similar to that of ordinary neural networks. The basic CNN architecture is a stack of convolutional layer (Conv), nonlinear layer (e.g. ReLU), pooling layer

(e.g. Max-pooling), and a loss function (e.g. SVM, Softmax) on the last fully connected (FC) (see Fig. 3.1). The output can be a single class (e.g. normal, benign, malignant) or a probability of classes that best describes the image. The input to a convolutional layer is a W1 $\times$ H1$\times$ D1 image where W1 is the width and H1 is the height of the image and D1 is the number of channels, e.g. an RGB image has D1=3. The convolutional layer will have F filters (e.g. 12 filters) of size N $\times$ N $\times$ D1 where N is smaller than the dimension of the image and D1 is the same as the number of channels (e.g. 5 $\times$ 5 $\times$ 3 (i.e. 5 pixels width and height, and 3 because images have depth 3, the color channels). During the convolution operation, each of the F filters convolves with the image to produce K feature maps of volume size W2 $\times$ H2 $\times$ D2 where: W2=H2=(W1-F+2P)/S+1, S is the number of strides, D2=F, and P is the number of zero paddings. For each feature map, a non-linear activation function is applied (e.g. ReLU). A non-linear activation function leaves the size of the volume unchanged (W2$\times$ H2 $\times$ D2). After applying ReLU, a down-sampling operation called Pool is applied along with the spatial dimensions (width, height) of the result feature map. After pooling, there may be any number of fully connected layers that compute the class scores (see Fig. 3.1). The CNN models are trained using back-propagation and gradient descent as for standard artificial neural networks. More details about the architecture of CNNs can be found in [9, 127, 207].

Convolutional Layer

In the convolutional layers, the activations from the previous layers are convolved with a set of small parameterized kernels (often called masks or filters), frequently of size 3 $\times$ 3 or 5 $\times$ 5. A different set of feature maps (k) within an image are extracted by sliding convolutional kernel on

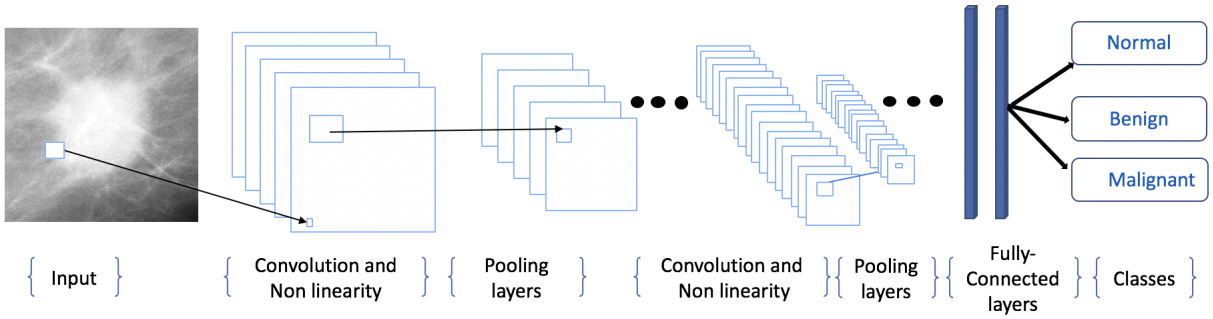


Fig. 3.1: The CNN architecture is a stack of Convolutional layer (Conv), Nonlinear layer (e.g.ReLU), Pooling layer (Pool), and a Loss function (e.g. SVM/Softmax) on the last (Fully connected) layer. The output can be a single class (e.g. Normal, Benign, Malignant).

the whole image with the same set of weights. By having each kernel share the exact same weights across the whole input domain, thus a drastic reduction in the number of weights that need to be learned. The motivation for this weight-sharing is that features such as edges appearing in one part of the image likely also appear in other parts. If you have a filter capable of detecting horizontal and vertical lines, say, then it can be used to detect them wherever they appear. Applying all the convolutional filters at all locations of the input to a convolutional layer produces a tensor of K feature maps.

Activation Layer

The K feature maps from a previous convolutional layer are fed through nonlinear activation functions. This makes it possible for the entire neural network to approximate almost any nonlinear function. The activation functions are generally the very simple rectified linear units or ReLUs. Feeding the feature maps through an activation function produces new tensors, typically also called feature maps. Recently, many variations of Rectified Linear Unit (ReLU) function have been pro-

posed [82] such as leaky ReLU, parametric ReLU, and randomized ReLU (see Fig. 3.2).

There are several other popular activation functions such as sigmoid, and Tanh. There are several advantages of using the ReLUs to Sigmoid/Tanh functions. Sigmoid presents a serious disadvantage called vanishing gradient problem. Sigmoid function's values are within the following range [0,1], and due to its nature, small and large values passed through the sigmoid function will become values close to zero and one respectively. This means that its gradient will be close to zero and learning will be slow. Slow learning is one of the things we want to avoid in DL since it results in expensive and tedious computations. The Tanh squashes a real-valued number to the range [-1, 1]. Like the sigmoid neuron, Tanh'activations saturate, but unlike the sigmoid neuron, its output is zero-centered [167].

ReLU became a popular choice in DL and even nowadays provides outstanding results. It came to solve the vanishing gradient problem mentioned before. ReLU has gradient 1 for positive inputs and 0 for negative ones (Equations 3.1, 3.2). The fact that the gradient is zero helps to make the network sparse keeping the useful links. Sparsity helps to keep the network less dense and decreases the computation, however, once the gradient is zero the corresponding nodes don't have any influence on the network anymore. This is a good idea since disconnecting some neurons may reduce overfitting. The ReLU activation function has the following form:

$$relu(x) = \begin{cases} 0, & \text{if } x <= 0 \\ x, & \text{if } x > 0 \end{cases} \tag{3.1}$$

which has the gradient:

$$\frac{d}{dx} relu(x) = \begin{cases} 0, & \text{if } x <= 0 \\ 1, & \text{if } x > 0 \end{cases} \tag{3.2}$$

Leaky ReLUs are one attempt to overcome the ”dying ReLU” problem [144]. Instead of the function being zero when x < 0, a leaky ReLU will instead have a small negative slope (α of 0.01, or so), see equations 3.3, 3.4. This reduces the sparsity but, on the other hand, makes the gradient more robust for optimization since now the weight will be adjusted for those nodes that were not active with ReLU. The Leaky ReLU activation function has the following form:

$$lrelu(x) = \begin{cases} \alpha x, & \text{if } x <= 0 \\ x, & \text{if } x > 0 \end{cases} \tag{3.3}$$

which has the gradient:

$$\frac{d}{dx} lrelu(x) = \begin{cases} \alpha, & \text{if } x <= 0 \\ 1, & \text{if } x > 0 \end{cases} \tag{3.4}$$

where α is a small constant as 0.01. When α is not 0.01 then it is called randomized ReLU.

A detailed explanation of the advantages and disadvantages of different activation functions are discussed in [82, 127, 167]. Theoretically leaky ReLU is better than ReLU. However, with a proper setting of the learning rate, and careful initialization of the weights, the ”dying ReLU” problem is less frequently an issue.

Pooling Layer

Once K feature maps are extracted, its exact location becomes less important as long as its approximate position relative to others is preserved. The pooling layer is also known as Sub-sampling or Down-sampling or Spatial-Pooling. It diminishes the dimensionality of each feature map but holds

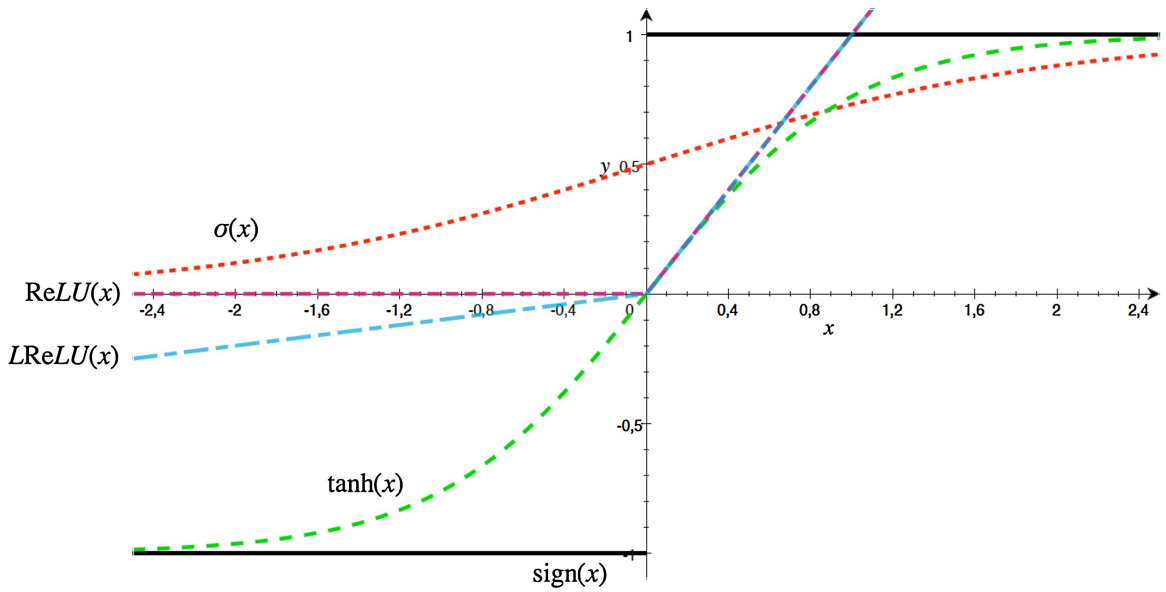


Fig. 3.2: Overview of classical (sign(x), σ(x)), and tanh(x)) and modern activation functions, like the Rectified Linear Unit (ReLU(x)) and the leaky ReLU (LReLU(x)) (source image [145])

the most critical data. Pooling operations take small grid regions as input and produce single numbers for each region. The number is usually computed by using the max function (max-pooling) or the average function (average pooling). For-example, in Max-Pooling, we select a spatial window (for instance, a 2×2 window) and take the biggest value output from the rectified feature map in that window. Rather than taking the biggest value, we could likewise take the average (Average-Pooling) in that window (see Fig. 3.3).

The use of pooling operation helps to extract a combination of features, which are invariant to transnational shifts and small distortions. Since a small shift of the input image results in small changes in the activation maps, thus, the pooling layers give the CNN some translational invariance. Reduction in the size of the feature map to invariant feature set not only regulates the complexity of the network but also helps in increasing the generalization by reducing overfitting. Generally, Max-pooling is used to make the input representations smaller and more manageable,

reduces the number of parameters and computations in the network, therefore, controlling overfitting. Different types of pooling formulations such as max, average, L2, overlapping, spatial pyramid pooling, etc. are used in recent CNNs. A different way of getting the down-sampling effect of pooling is to use convolutions with increased stride lengths.

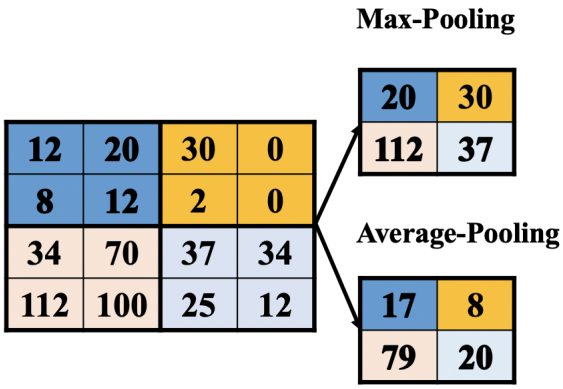


Fig. 3.3: Examples of 2×2 pooling, max-pooling or average-pooling that reduces the data dimension by half.

Dropout regularization [127]

Dropout regularization is a simple idea that gave a huge boost in the performance of CNNs. Dropout randomly kills neurons during training, which ends up using slightly different networks for each batch of training data, and the weights of the trained network are tuned based on optimization of multiple variations of the network (see Fig. 3.4). This means that their (killed neurons) contribution to the activation of downstream neurons is temporally removed on the forward pass and any weight updates are not applied to these neurons on the backward pass. Dropout is usually only applied after fully connected layers, but not after convolutional layers.

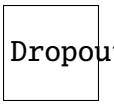


Fig. 3.4: (a) shows a standard network without dropout. (b) shows a network after dropout some neurons during training.

Batch normalization [105]

Batch normalization (BN) layers are placed after the activation layers, producing normalized activation maps by subtracting the mean and dividing by the standard deviation for each training batch. Including batch normalization layers forces the network to periodically change its activations to zero mean and unit standard deviation as the training batch hits these layers, which works as a regularizer for the network, speeds up training, and makes it less dependent on careful parameter initialization.

Fully Connected Layer

Neurons in this layer are fully connected to all neurons in the previous layer, as in a regular Neural Network.

Loss Layer

The last fully-connected layer serves as the loss layer that computes the loss or error which is a penalty for the discrepancy between desired and actual output. For predicting a single class out of K mutually exclusive classes Softmax loss is used. It maps the predictions to non-negative values and also normalized to get probability distribution over classes. One benefit of using Softmax output is that the values of all output units. so the output can be interpreted as a probability distribution

over all classes under a classification problem domain.

With methods like the ReLU, BN, Dropouts and deep residual learning alleviating issues such as the vanishing gradient problem, deeper models can be trained more efficiently and hence pushing deep learning to another level.

3.2 Popular CNNs

CNN history begins from the neuro-biological experiments conducted by Hubel and Wiesel (1959, 1962) [101]. Their work provided a platform for many cognitive models, almost all of which were latterly replaced by CNN. Over the last few years, different efforts have been carried out to improve the performance of CNNs [92, 127, 130, 201, 215, 248]. The advancements in CNNs can be categorized in different ways including activation, loss function, optimization, regularization, learning algorithms, and restructuring of processing units.

LeNet [130], Alex-Net [127], ZF-Net [248], GoogLeNet [215], VGG-Net [201] and ResNet [92] have been extensively used as a pre-trained networks to classify images for medical domains instead of training a network from scratch. Table 3.1 shows the configurations of the most popular CNNs. Generally, training a deep CNN requires extensive computational and memory resources. Training these networks from scratch typically takes days or weeks on modern GPUs (Table 3.1). All these networks were trained on the 1000 object category classification on the ImageNet data-set [187]. The ImageNet data-set consists of a 1.2M image training set, a 50K image validation set, and a 100K image test set. Two error rates are reported for these networks: top-1 and top-5, where the top-5 error rate is the fraction of test images for which the correct label is not among the five labels considered most probable by the model. All these network architectures use the data

augmentation technique to prevent overfitting with dropout initially set to 0.5. Table 3.1 shows that the number of layers is going deeper and deeper within the newer implementations as in ResNet. In the following subsections, we will give a brief introduction to the top CNN architectures and a summary of their main contributions.

Table 3.1: The configurations of AlexNet, ZF-Net, GoogLeNet, VGG-Net, and ResNet models.

	AlexNet [127]	**ZF-Net** [248]	**GoogLeNet** [215]	**VGG-Net** [201]	**ResNet** [92]
Year	2012	2013	2014	2014	2015
Image Resolution	227×227	227×227	224×224	224×224	2244×224
Number of layers	8	8	22	19	152
Number of Conv-Pool layers	5	5	21	16	151
Number of FC layers	3	3	1	3	1
Full connected layer size	4096,4096,1000	4096,4096,1000	1000	4096,4096,1000	1000
Filter Sizes	3, 5, 11	3, 5, 11	1,3,5,7	3	1,3,7
Number of Filters	96 - 384	96 - 384	64 - 384	64 - 512	64 - 2048
Strides	1, 4	1, 4	1, 2	1	1, 2
Data Augmentation	+	+	+	+	+
Dropout	+	+	+	+	+
Batch Normalization	-	-	-	-	+
Number of GPU	2 GTX 580 GPUs	1 GTX 580 GPUs	A few high-end GPUs	4 Nvidia Titan Black GPUs	8 GPUs
Training Time	5:6 days	12 days	1 week	2:3 weeks	2:3 weeks
Top-5 error	16.40%	11.2%	6.70%	7.30%	3.57%

3.2.1 LeNet

LeNet [130] was one of the first pioneering work in CNNs by LeCun et al [130]. In their work, the task of Handwritten Digit Recognition was performed using CNNs (see Fig. 3.5). The lack of high computing machines at that time led to a break in the use of CNN. LeNet is a feed-forward NN that constitutes of five alternating layers of convolutional and pooling, followed by two fully connected layers. LeNet exploited the underlying basis of image that the neighboring pixels are correlated to each other and are distributed across the entire image. Therefore, convolution with learnable parameters is an effective way to extract similar features at multiple locations with few parameters.

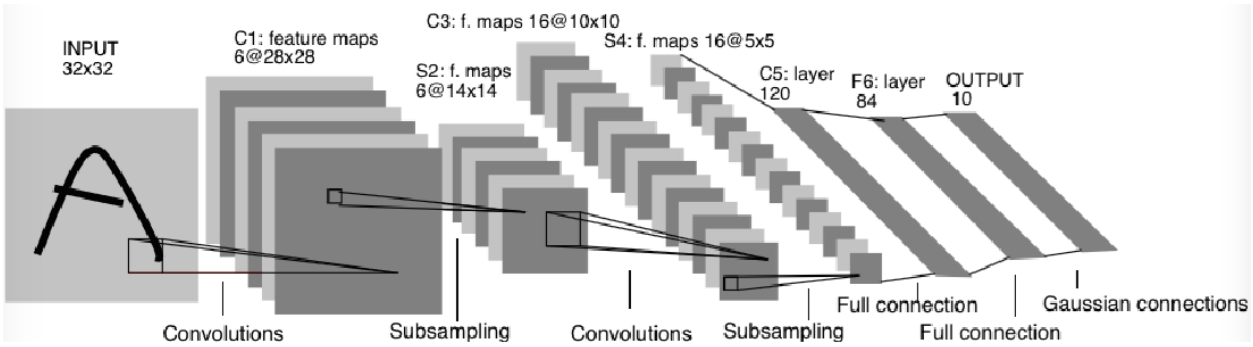


Fig. 3.5: Architecture of LeNet (source image from [130]). Two convolutional layers and two sub-sampling layers are interleaved to form the first four layers. Activation layers are appended after every layer up. Two FC layers are attached after the last sub-sampling layers to vectorize image representations. The last and output layer is composed of Euclidean Radial Basis Function (RBF) units which output the Euclidean distance between the network outputs and ground truth labels for ten classes.

3.2.2 Alex-Net

Alex-Net [127] was the first CNN to win the ImageNet Challenge in 2012. AlexNet is considered as the first deep CNN architecture, which showed groundbreaking results for image classification and recognition tasks. AlexNet's CNN consists of five Conv layers and three fully connected (FC) layers. Within each Conv layer, there are 96 to 384 filters and the filter size 3×3, 5×5, 11×11, with 3 to 256 channels each (see Fig. 3.6). A ReLU, non-linearity function, is used in each layer. Max-pooling of 3×3 is applied to the outputs of layers 1, 2 and 5. Alex-Net used a stride of 4 in the first layer of the network. AlexNet's model requires 61M weights to process one 227×227 input image (top-5 error of 16.40%). Krizhevesky et al. [127] used large size filters (11×11 and 5×5) at the initial layers, compared to previously proposed networks. Krizhevesky et al. enhanced the learning capacity of the AlexNet CNN by making it deeper and by applying a number of parameter optimizations strategies. To address the problem of data overfitting which is caused by increasing the depth of the model, Krizhevesky et al. [127] introduced the dropouts algorithm which randomly skips some transformational units during training to enforce the model to learn more robust features. In addition to this, ReLU was employed as a non-saturating activation function to improve the convergence rate by alleviating the problem of vanishing gradient. AlexNet was trained in parallel on two NVIDIA GTX 580 GPUs to overcome shortcomings of the hardware at that time. AlexNet has started a new era of research in the architectural advancements of CNNs.

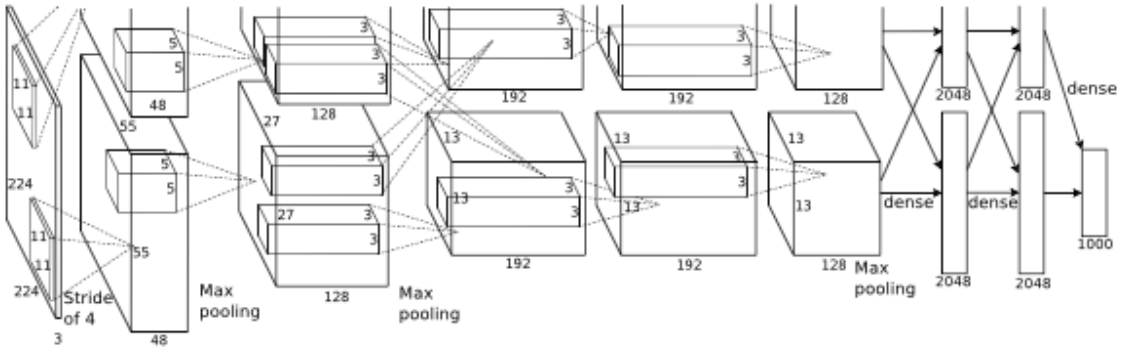


Fig. 3.6: Architecture of AlexNet (source image from [127]). AlexNet consists of five convolutional layers and three sub-sampling (shown as max-pooling) layers are interleaved to form the first eight layers. Activation layers are appended after convolutional layers. Two fully connected layers are attached after the last sub-sampling layers to vectorize image representations. The last and output layer outputs the softmax loss of the network predictions for 1000 classes.

3.2.3 ZF-Net

Before 2013, CNN models were largely based on hit-and-trial, without knowing the exact reason behind the improvement. This lack of understanding limited the performance of deep CNNs on complex images. In 2013, Zeiler and Fergus [248] proposed a multilayer De-convolutional NN, which got famous as ZF-Net. ZF-Net is a slightly modified version of the Alex-Net model and uses an interesting way of visualizing their feature maps (see Fig. 3.7). In ZF-Net, the used visualization technique gives insight into the function of intermediate feature layers and the operation of the used classifier. The objective of ZF-Net was to monitor the learning rate during training and thus enhance the Alex-Net model. Their experiments on Alex-Net showed that only a few neurons were active during the training process, while other neurons were dead in the first and second layers of the network. Based on that, Zeiler and Fergus adjusted the Alex-Net architecture and performed

parameter optimization. Zeiler and Fergus maximized the learning of CNN by reducing both the filter size and stride to retain a maximum number of features in the first two convolutional layers. This readjustment in CNN architecture boosted the performance, which suggested that features visualization can be used for the identification of design shortcomings and for timely adjustment of parameters.

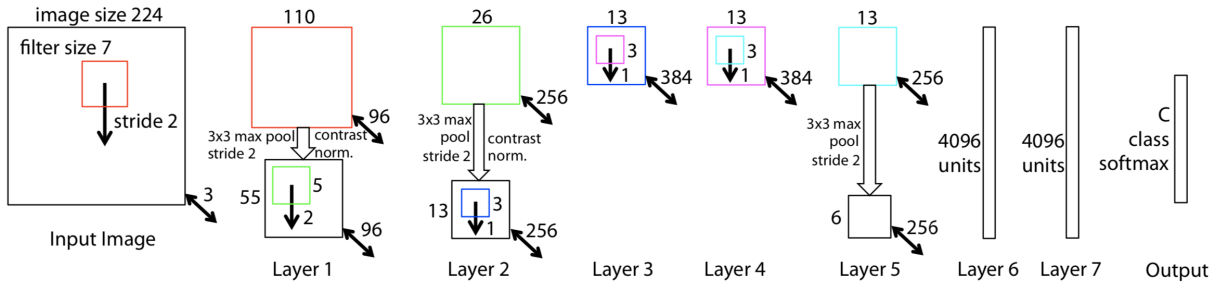


Fig. 3.7: Architecture of ZF-Net (source image from [248]). The ZF-Net architecture consists of 8 layers. A 224 × 224 crop of an image (with 3 color planes) is presented as the input. This is convolved with 96 different 1st layer filters (red), each of size 7 × 7, using a stride of 2 in both x and y. The resulting feature maps are then: (i) passed through a rectified linear function, (ii) max-pooled within 3 × 3 regions, using stride 2 and (iii) contrast normalized across feature maps to give 96 different 5 × 5 element feature maps. Similar operations are repeated in layers 2, 3, 4, 5. The last two layers are FC layers, taking features from the top convolutional layer as input in vector form.

3.2.4 The VGG-Net

Simonyan et al. [201] proposed a simple and effective CNN model called VGG-Net. The VGG-Net model reinforces that the CNNs have to have a deep network of layers. VGG-Net was made 19 layers deep compared to AlexNet and ZF-Net which had a few numbers of layers (see Fig. 3.8).

The VGG-Net won the 2013-ILSVRC competition. Simonyan et al. [201] suggested that small size filters can improve the performance of the CNNs. Based on these findings, VGG-Net replaced the 11×11 and 5×5 filters with a stack of 3×3 filters layer and experimentally demonstrated that concurrent placement of 3×3 filters can induce the effect of the large size filter. The use of the small size filters provides an additional benefit of low computational complexity by reducing the number of parameters. These findings set a new trend in research to work with smaller size filters in CNN. VGG-Net regulates the complexity of the network by placing 1×1 convolution in between the convolutional layers, which also, learn a linear combination of the resultant feature maps. For the tuning of the network, max pooling is placed after the convolutional layer, while padding was performed to maintain the spatial resolution. VGG-Net suffered from a high computational burden due to the use of about 140 million parameters.

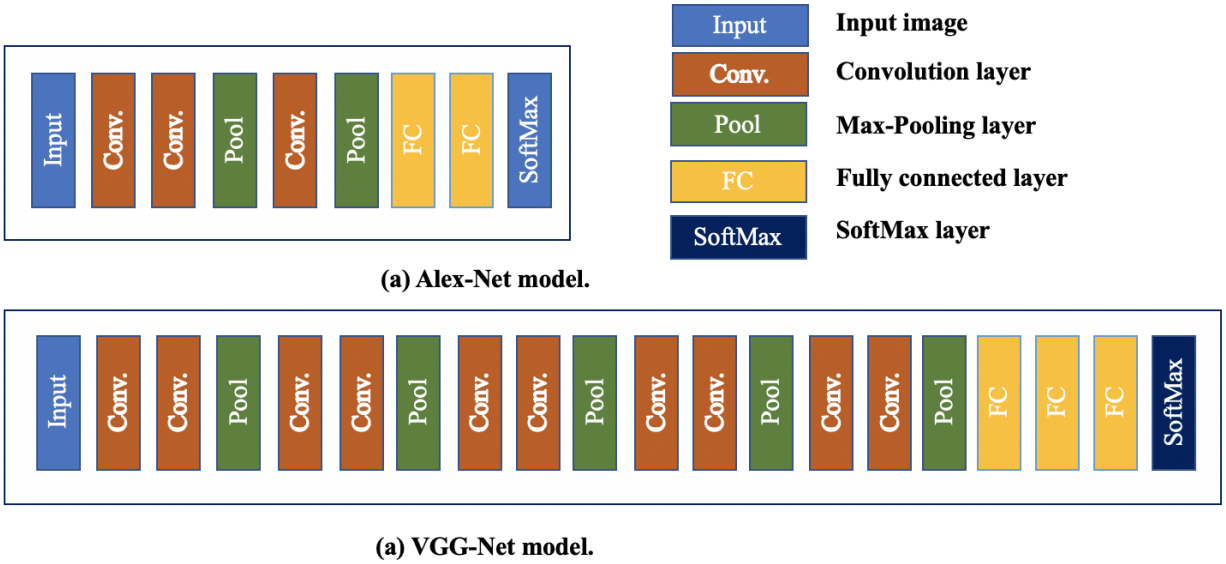


Fig. 3.8: Architecture of VGG-Net in comparison to AlexNet (source image [116])

3.2.5 GoogLeNet

GoogLeNet model [215] was the winner of the 2014-ILSVRC competition. GoogleNet has 22 layers. It introduced an inception module to the CNN model (Figs 3.9 and 3.10). It has pieces of the network that are working in parallel in contrast to previous CNN models, which have only a single serial connection. It introduced the concept of inception block in CNN, whereby it incorporates multi-scale convolutional transformations using split, transform, and merge ideas. The architecture of inception block. This block encapsulates filters of different sizes (1×1, 3×3, and 5×5) to capture spatial information at fine and coarse scales. GoogleNet regulates the computation by adding a 1×1 convolutional filter, before employing large size kernels. It used sparse connections, to overcome the problem of redundant information and reduced cost by omitting feature maps that were not relevant. Furthermore, connection's density was reduced by using global average pooling at the last layer, instead of using an FC layer. These parameter optimizations caused a significant decrease in the number of parameters from about 40 million to 5 million parameters.

3.2.6 ResNet

ResNet (Residual Net) [92] was proposed by He et al., which is considered as a continuation of deep Nets [92]. ResNet uses residual connections to go even deeper. ResNet revolutionized the CNN architecture by introducing the concept of residual learning in CNN (see Fig. 3.11). The implementation of a residual block [92], is straightforward: for every few convolutional layers, a shortcut connection is added that runs parallel to these layers and implements the identity mapping. The output of the convolutional layers is then added to the output of the shortcut branch and the result is propagated to the subsequent block (see Fig. 3.11). Besides the use of shortcut connec-

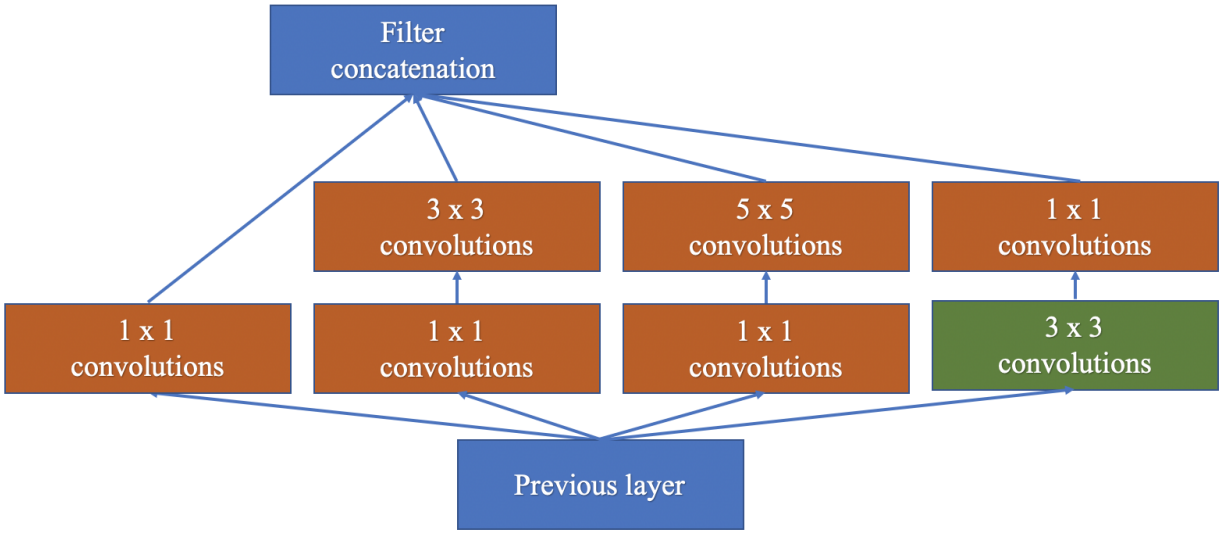


Fig. 3.9: Inception layer of GoogLeNet. Two 1×1 convolutions stacked before expensive 3×3 and 5×5 convolutions are used to reduce dimensions, and also combined with ReLU activation.

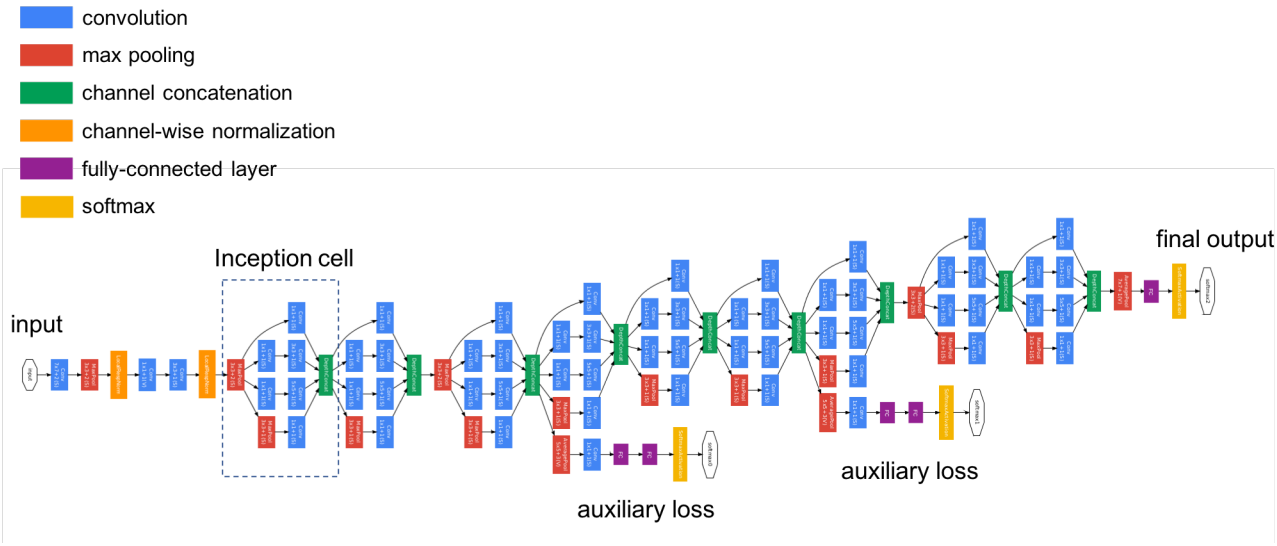


Fig. 3.10: Architecture of GoogleNet model (source [215]).

tions, network architecture is mainly inspired by the philosophy of VGG16-Net and VGG19-Net. All convolutional layers have small kernels of size 3×3. ResNet has two simple design rules: (i) for the same output feature map size, the layers have the same number of filters; (ii) when the

feature map size is halved (with convolutional layers of stride 2), the number of filters is doubled to preserve the time complexity per layer. The authors tested architectures of varying depth in the range between 34 and 152 layers. ResNet, which is 20 and 8× deeper than AlexNet and VGG respectively, showed less computational complexity than previously proposed CNNs. The ResNet with 152 layers was the winner of the ImageNet challenge 2015 [187] (top-5 error of 3.57%). it has 60M parameters.

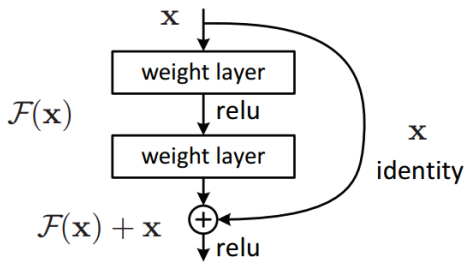


Fig. 3.11: Architecture of the basic building block of ResNet is a Residual block which is repeated through out the network [92].

3.2.7 DenseNets

DenseNet was proposed to solve the vanishing gradient problem [99]. DenseNet connects each layer to every other layer in a feed-forward fashion, thus feature maps of all preceding layers were used as inputs into all subsequent layers. DenseNet concatenates the features of the previous layer instead of adding them, thus, the network may gain the ability to explicitly differentiate between information that is added to the network and information that is preserved. DenseNet has a narrow layer structure; however, it becomes parametrically expensive with an increase in several feature maps.

3.2.8 Attention Based CNNs

Wang et al. [224]proposed a Residual Attention Network to improve feature representation of the network. The motivation behind the incorporation of attention in CNN was to make a network capable of learning object aware features. Attention-based CNNs capture different levels of abstraction and focus on features relevant to the context which play a significant role in image localization and recognition.

3.2.9 YOLO

YOLO [181]is another famous CNN named that is recently used for object classification and localization while processing the image only once, as is implied by its name, You Only Look Once. YOLO uses a single CNN operating directly on the image and outputting bounding boxes and class probabilities. It incorporates several elements from the above networks, including inception modules and pre-training a smaller version of the network. It's fast enough to enable real-time processing.

3.2.10 GANS

A generative adversarial network [249] consists of two neural networks compete against each other. The generative network G is tasked with creating samples that the discriminative network D is supposed to classify as coming from the generative network (fake) or the training data (real). The networks are trained simultaneously, where G aims to maximize the probability that D makes a mistake while D aims for high classification accuracy.

3.2.11 U-Net

U-Net [186] is a popular network for segmentation in 2D images. It is a fully convolutional network that down-samples the input image through am encoder CNN, before being up-sampled using transpose convolutions in a decoder until it reaches its original size. Also, there are skip connections that concatenates features from the down-sampling to the upsampling paths (see Fig. 3.12).

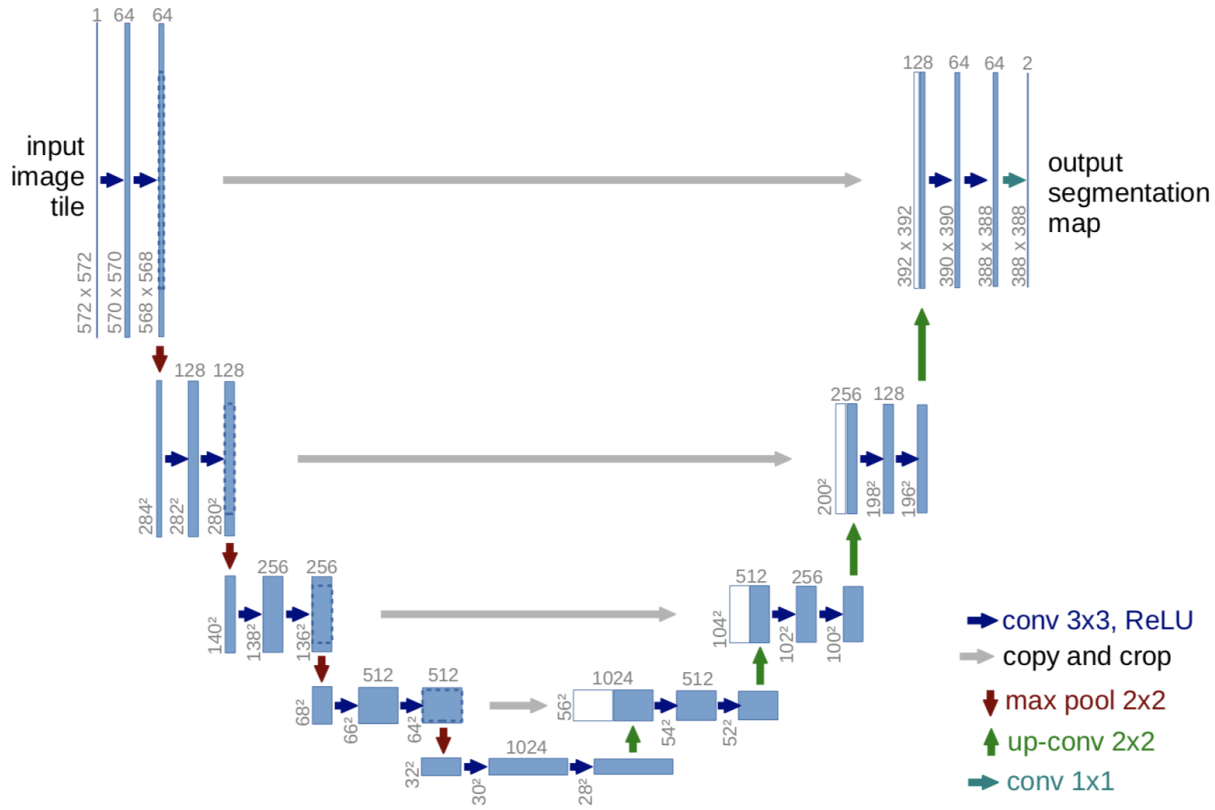


Fig. 3.12: Original U-Net architecture (source image from [186]. Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower-left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

3.2.12 Convolutional Residual Networks (CRNs)

Deeper networks have a higher capability to learn as in the VGG-Net, but deeper networks not only suffer from gradient vanishing problem but also face degradation. It means with the depth increasing, the accuracy gets saturated and then rapidly degrades. To take advantage of deeper network structure, He et al. [92] introduced the residual networks which were initially developed for natural image segmentation on 2D images. In this model, the residual block is introduced, instead of consecutively feeding the stacked layers with the feature map, a residual map is fed to every few layers. In other words, the residual maps are skipped connections, allowing the network to redirect the derivatives through the network by skipping some layers. This design helped the network to enjoy the accuracy gained from deeper designs (see Fig. 3.13).

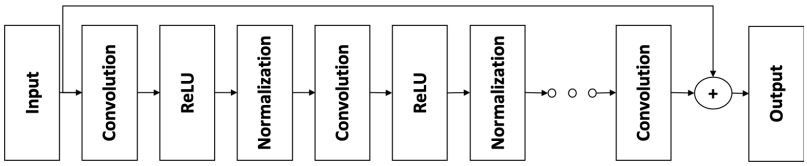


Fig. 3.13: A residual block of CRN. Residual block may have various number and combination of convolutional layers inside, depending on the network design and depth.

3.3 Performance Metrics

Researchers in the medical field use the following performance validation metrics to evaluate their method: sensitivity (SEN), specificity (SPE), accuracy (ACC), ROC curve, and F1 metric.

The sensitivity (SEN), recall, hit rate, or true positive rate (TPR) represents the proportion of positive cases correctly classified (Equation 3.5). Where TP is true positive, FN is false negative,

TN is true negative, and FP is false positive.

$$SEN = \frac{TP}{TP + FN}, \tag{3.5}$$

Precision (also called positive predictive value (PPV)) is the fraction of positive cases among the retrieved instances (Equation 3.6).

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{3.6}$$

The specificity (SPE), selectivity or true negative rate (TNR) is the proportion of actual negative cases which are correctly classified (Equation 3.7).

$$SPE = \frac{TN}{TN + FP}, \tag{3.7}$$

The accuracy (ACC) represents the proportion of true results (both true positives and true negatives) (Equation 3.8).

$$ACC = \frac{TP + TN}{TN + TP + FN + FP}, \tag{3.8}$$

The Receiver operating characteristic (ROC) curve is the plot between sensitivity and (1- specificity). (1- specificity) is also known as false positive rate (FPR).

The AUC measure is computed just by obtaining the area under of ROC curve (Equation 3.9).

$$AUC = \frac{1 + TPR - FPR}{2}, \tag{3.9}$$

The F1 score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0, (Equation 3.10).

$$F1_{score} = \frac{2 \times precision \times recall}{precision + recall}, \tag{3.10}$$

The accuracy is not a useful measure when evaluating classifiers learned on imbalanced data-sets [46]. The F-value combines the trade-offs of precision and recall and outputs a single number reflecting the goodness of a classifier in the presence of imbalanced data-sets.

Chapter 4

Convolution Neural Networks Best Practices for Mammography

In this chapter, we explain the practices that contribute to improving the performance of CNNs for MGs. We are going to highlight and focus on some of them that show significant changes in the classification accuracy when applied to MG images. Recent survey papers [129, 139, 150] discuss more trends for natural images.

4.1 Data Preparation

4.1.1 Pre-Processing of MG Images

Pre-processing of MG images is an essential task before training CNNs [26, 106, 111, 114, 192]. The pre-processing consists of contrast enhancement, noise removal, and breast segmentation. Breast segmentation includes the removal of the background area, labels, artifacts, and pectoral muscle which disturb the detection of Mass/MCs [17, 71]. It is important to have good separation between foreground and background pixels and do not remove the important information in images [30, 194, 195]. The commonly used filters for image enhancement and noise reduction are the adaptive mean filter, median filter, and contrast limited adaptive histogram equalization (CLAHE) [4, 5, 36, 117, 218].

4.1.2 Image Size, Cropping, and Down-Sampling

Most studies have used segmented ROIs to reduce the computation of the CNNs and to avoid the issue of small training data. These ROIs can be obtained by manual segmentation of the images using the available ground truth data, or an automatic detection system. The ROIs are cropped and re-scaled to r $\times$ r pixels with the lesion centered within the image. However, using very small subsampled (e.g. 32×32) patches may not contain enough detail to improve the classification results as in [44, 59, 76, 95, 106, 154, 188, 192, 236].

Two strategies have been utilized to use full image size for training CNNs on MGs instead of ROIs. The first strategy, down-sample high-resolution images to $\approx 250 \times 250$. However, the requirement to find small mass regions or MCs clusters in down-sampled high-resolution images is unlikely to be successful for MGs [143]. The second strategy, train a patch-level CNN classifier, which is then used as a feature extractor to an image-level model. In the image-level model, each image is partitioned into a set of patches with a minimal overlap such that each patch is contained entirely within the image. The final classification involves aggregation across patches and the CC & MLO views [143].

4.1.3 Mixing Databases

In literature, researchers mix several databases to analyze their CNNs. The fusion from different image types (FSM and FFDM) assists CNNs in terms of detection rate. Researchers in [3, 40, 41, 55, 104, 190, 258] compared both image quality and detection on FFDM and FSM databases. They have shown that CNN using FFDM images gives a better detection rate than using FSM images. Moreover, these studies show that DL training using the fusion of both FFDM and FSM lower the

number of false detections [76, 190].

4.1.4 Learned and Hand-Crafted Features

The hand-crafted features (i.e. Haar-like features, a histogram of oriented gradients (HOG), and histogram of the gradient divergence (HGD) are commonly used with traditional machine learning approaches for object recognition like support vector machines. CNNs can extract features from the input image data-sets. Thus, CNNs remove the necessity of the time-consuming hand-crafted features.

However, the authors in [24, 26, 43, 56, 57, 121, 124, 176, 194, 244] have demonstrated the importance of combining the extracted features using deep CNNs with hand-crafted features like texture, and shape. Interestingly, the combination of both representations (learned and hand-crafted features) resulted in a better descriptor for Mass/MCs lesion classification [24, 26]. The reason behind using hand-crafted features is that the learning process should be guided by a training data-set that has a wide variability of texture and shape features. For example, Dhungel et al. [56] proposed a two-step training process involving pre-training based on a large set of hand-crafted features. The second stage fine-tunes the features learned in the first stage to becoming more specialized for the classification problem.

Using hand-crafted features depends on the size of the data-set. With a small training data-set, generating hand-crafted images could result in a better model for Mass/MCs lesion classification. Also, employing some hand-crafted features that specifically target small and missed lesions is a more effective strategy than adding extra cases to the train a data-set. Thus, the performance of CNNs trained with small data-set can be improved by incorporating hand-specified features to deal

with cases that cause false positives or false negatives [244].

4.2 Hyper-Parameters

Hyper-parameters are variables that determine the network structure (e.g. number of hidden layers), and the variables which determine how the network is trained (e.g. learning rate). Hyper-parameters are manually chosen before training the CNNs.

4.2.1 Data Augmentation

Data augmentation is an appealing solution to reduce overfitting and increase the generalization of the model and boost performance. Overfitting happens in CNNs when the models learn too well the details from training data, but they do not generalize well from the training data, to make good predictions unseen data. As a result, the performance of the trained model is poor for testing data. That usually happens when the size of the training data-set is too small compared with the number of model parameters that need to be learned.

Data augmentation artificially creates new sample images by applying transformations like flipping, and rotation to the actual data. Common data augmentation techniques for mammography images are horizontally flipping, rotations (90, 180, and 270 degrees), jittering, and random scaling. Such data augmentation generates relevant training samples because tumors may present in various orientations and sizes. Thus, augmentation techniques do not change the underlying pathology of the masses. Data augmentation has been employed by many studies [13, 15, 17, 26, 31, 43, 44, 48, 56, 62, 65, 71, 75, 79, 87, 106, 108, 111, 112, 122–124, 135, 140, 151, 154, 174, 188, 189, 192, 211, 228, 236, 238, 244, 258].

4.2.2 Going Deeper

In CNN, the design of the network architecture completely depends on the model requirements and the size of the data-set. The CNNs in [65, 106, 124] have a fewer number of layers but show good accuracy. However, the work done in [48, 111, 192, 215] shows that we can get better performance in term of the higher area under the ROC curve (AUC) as the architecture goes deeper and trained on more data. Deep architectures can lead to abstract representations because more abstract shapes can often be constructed in terms of less abstract ones captured in earlier layers. Adding more layers will help the model to extract more features. But adding more layers can be done to a certain extent and there is a limit. After that, instead of extracting features, it results in overfitting the network that can lead to false positives. Adding more hidden layers will promote the accuracy of large data-sets. Adding layers unnecessarily to a CNN will increase the number of parameters, and for a smaller data-set, it will reduce the accuracy of the test data. Deep architectures are often challenging to train effectively, and this has been the subject of more recent research. Choosing a smaller network or a larger one cannot be estimated theoretically. A trade-off between accuracy and deep networks needs to be done with trial and error method and some experience and practice based on the data-set.

4.2.3 Learning Rate

Learning rate (LR) is one of the most important hyper-parameters, which influences the CNNs' performance. Deep learning models are typically trained by a stochastic gradient descent optimizer. There are many variations of stochastic gradient descent as Adam, RMS Prop, Adagrad, etc. All these optimizers let users set the learning rate. The learning rate controls how much the

network parameters are adjusted to minimize the network's loss function. If the LR is too small, the CNN will converge after many iterations to the best values. However, if LR is too high, it can cause undesirable divergent behavior in the loss function. Famous learning rate policies are step decay, quadratic decay, square root decay, and linear decay [150]. A common practice when dealing with MG images, is to use a step decay rate where the LR is reduced by some percentage after a set number of training epochs. For example, Yi et al. [244] used a learning rate of 0.001 with a decay rate of 0.99 per epoch, and a regularization coefficient of 10^{-5} for training their CNN. Another common practice is to use a small learning rate (e.g. 0.001) to train a pre-trained network since we expect well-adjusted pre-trained weights compared to randomly initialized weights.

4.2.4 Activation Functions

Recently, many variations of rectified linear unit (ReLU) function have been proposed for activation function such as leaky ReLU, parametric ReLU, and randomized ReLU [82]. There are other popular activation functions such as Sigmoid, and Tanh. The activation functions bring non-linearity into CNNs. Sigmoid presents a serious disadvantage called the vanishing gradient problem. In the vanishing gradient problem, the gradient of small input values to sigmoid functions tends to get smaller (close to zero) as gradients are computed backward through the hidden layers, resulting in slow learning in the earlier layers of the model. Slow learning is highly avoided in DL since it results in expensive and tedious computations [167].

ReLU became a popular choice in DL and even nowadays provides outstanding results as it solves the vanishing gradient problem [82]. ReLU has a gradient one for positive inputs and zeroes for negative inputs. As long as values are above zero, the gradient of the activation function will be

one, meaning that it can learn anyways. This solves the vanishing gradient problem present in the sigmoid activation function. On the downside, once the gradient is zero the corresponding nodes do not have any influence on the network anymore, which is known as the "dying ReLU" problem. Leaky ReLU is one attempt to overcome the dying ReLU problem [144]. Instead of the output of ReLU being zero when input is less than zero, a leaky ReLU will provide a small negative slope (α of 0.01, or so). This small slop reduces the sparsity but, on the other hand, makes the gradient more robust for optimization, since in this case, the weight will be adjusted for those nodes that were not active with ReLU. When the slop is not constant (e.g. 0.01) then it is called randomized ReLU.

A detailed explanation of the advantages and disadvantages of different activation functions are discussed in [82, 127, 167]. Theoretically, leaky ReLU is in general better than ReLU. However, ReLU has been chosen as an activation function in most of the CNNs for MGs as it allows faster learning [113, 135, 143, 153, 173, 236].

4.3 Techniques for Improving the CNNs Performance

4.3.1 Dropout

Dropout is a regularization technique proposed in [208] that superior the other regularization methods (L1, L2, Max norm). Dropout prevents a CNN model from overfitting. This technique randomly selects neurons and ignore them during training (Fig. 3.4). They are "dropped-out" randomly. This means that their contribution to the activation of downstream neurons is temporally removed on the forward pass and any weight updates are not applied to these neurons on the backward pass [127]. Smirnov [203] has shown a comparison of regularization methods with deep

CNNs and showed that the dropout technique is in general better than other regularization techniques. The authors in [48, 58, 79, 122, 123, 135, 154, 189, 189, 190, 236] have used dropout in their work with MGs. The dropout of 0.5 is a common value for mammography images.

4.3.2 Batch Normalization

In a CNN model, a batch normalization (BN) layer normalizes input variables across a mini-batch (a subset of the training data-set). First, the BN layer normalizes the activations of each channel by subtracting the mini-batch mean and dividing by the mini-batch standard deviation. Then, the BN layer shifts the input by a learnable offset β and scales it by a learnable scale factor γ, thus reduces the networks' internal covariant shift. BN speeds up the training of CNNs and reduces the sensitivity to network initialization. According to [105], BN allows the use of much higher learning rates and less care about initialization as it acts as a regularizer. BN results in faster convergence and as a consequence overall faster training for a CNN. Besides that, BN regulates the values going into each activation function. With BN, saturating nonlinear activation functions (e.g. sigmoid) that do not work well in deep networks tend to become viable again. Similar to dropout, BN adds some noise to each hidden layer's activations. Therefore, using BN causes less dropout value. BN has been used in CNNs for MG images [48, 143, 228]. For mammography, it is recommended to not depend only on BN for regularization; and to use it together with dropout.

4.3.3 Transfer Learning

Training a deep CNN requires large amounts of labeled training data [94]. Only few studies train an entire CNNs from scratch with random initialization; and the rest use TL approaches either fine-tune a pre-trained network [42, 48, 55, 65, 103, 111, 135, 151, 190, 192, 213, 214] or use a pre-

trained network as feature extractor [40, 103, 230, 236]. Recent overviews of TL in deep network models are given in [17, 103, 143, 155]. The need for TL in medical domain occurs because data are scarce and expensive, they are not publicly available, and it is time-consuming to collect and label them by professional radiologists [84, 103, 112, 162, 165, 216, 245]. Moreover, training a deep CNN requires extensive computational and memory resources [84, 127, 215].

References [214, 217, 248] show that the main power of a CNN lies in its deep architecture. Extracted features of earlier layers of a pre-trained CNN (i.e. on natural images) contain more generic features (e.g. edge detectors or blob detectors) that are useful for many tasks; but in later layers, generic features are combined and become more specific to the details of the classes contained in the training data-set. Thus, a deep CNN allows extracting a set of discriminating features at multiple levels of abstraction which can be transferable from one domain to another. However, the required level of fine-tuning differs from one application to another. Tajbakhsh et al. [217] show that neither shallow tuning nor deep tuning may be the optimal choice for a particular application. Moreover, layer-wise fine-tuning may offer a practical way to reach the best performance for a certain application and should be chosen experimentally. Also, the work in [87, 121, 189] has achieved a good performance on a small data-set by pre-training the network on a large data-set of general medical images.

Most of the studies employed TL have used ImageNet' data-set [187] for pre-training their network [41, 103, 111, 135, 151, 190, 214, 233, 240, 251]. The commonly used pre-trained CNNs architectures for mammography are Alex-Net [41, 71, 103, 111, 135, 151, 190, 214, 240, 251], VGG16 [71, 196, 240], ResNet50 [196, 240] and GoogLeNet [111, 135, 240]. All the deep CNN architectures that are pre-trained using ImageNet are designed for a 1000-class classification task. To adapt them

to the task at hand, the last three layers are removed from each network and three new layers (FC layer, soft-max layer, and classification layer) are appended to the remaining structure of each network.

Until large-scale medical image data-sets for mammography became available, the combination of TL and data augmentation is a very promising approach for training deep CNNs. By visualizing the features learned at different layers during the training process, a model can be monitored to closely observe and track its performance [244]. Learned features can indicate whether a model is successfully learning or not, allowing a user to stop the training process early [97].

4.3.4 Cross-Validation

Cross-validation is a statistical technique to evaluate predictive models by partitioning the original samples into a training set to train the model, and a test set to evaluate it. There are three common types used in literature for validation, the hold-out splits [132, 232], three-way data splits [79, 124, 135, 139, 143], the K-fold cross-validation [3, 58, 70, 151, 153, 175, 190, 220, 244, 256]. In the hold-out data splits, data is split into training set and test set (e.g. 80%, and 20%, respectively). The training set is used to train the model and the test set is used to estimate the error rate of the trained model. In the three-way data splits, data are randomly split into training, validation and testing sets. The CNN model is trained on the training set and is evaluated on the validation set. Training and validation may be iterated a few times until the best model is found. The final model is assessed using the test set.

In the K-folds cross-validation, data are split into k different subsets (or folds). The cross-validation process is repeated K times (the folds), with each of the K sets used exactly once as the

test set. The K error estimates from the folds can then be averaged to produce a single estimation. Cross-validation avoids overfitting and gives a less biased estimate of the performance of the model [59, 242]. In practice, the choice of the number of folds depends on the size of the data-set. In literature, the common strategy is to use K-fold cross-validation for mammography. For large data-sets, it is a common choice to use 3 to 5-fold cross-validation. For a small mammography data-set, it is a common choice to use 10-fold cross-validation.

4.3.5 Context and Patient Information

Integrating some information such as patient age, breast density and another context like the view type (CC or MLO) into a CNN method can improve the detection rate of CNNs [124]. Multi-modal machine learning aims to build models that can process and relate information from multiple modalities (e.g. images and text) with a score level fusion at the final prediction results.

4.3.6 Multi-View and Single-View Images

It is a good practice to use both CC & MLO views to detect abnormalities. A true abnormality can usually be detected on two different views of an MG. Recent studies in [33, 41, 58, 79, 121, 230, 238, 244, 256] lead to significant improvements of multi-view (MV) approach compared to single-view (SV) ones, demonstrating that the high-level features of the individual CNN models provide a robust representation of the input images. Comparing two views can aid in the reduction of false positives and false negatives.

4.3.7 Balanced and Imbalance Distribution

A couple of publicly available databases (e.g. INbreast, DDSM) is constructed to include approximately the same proportions of normal and abnormal cases, which is a balanced distribution of classes. Other databases called imbalance distribution (natural distribution) databases, which include unequal proportions of normal and abnormal cases. Training CNN models directly on imbalanced data-sets may bias the prediction towards the more common classes like normal, resulting in false negatives. Whereas the minority ones are misclassified frequently [146]. The authors in [17, 32, 40, 48, 65, 95, 103, 121, 135, 256] have pointed out that the balance of the number of samples per class has a great impact on the performance of the system. However, the authors in [79, 124, 154] used a natural distribution databases. According to [46] choosing a wrong distribution or objective function while developing a classification model can introduce a bias towards potentially uninteresting class (non-cancerous).

For MG images, it is preferable to use a balanced data-set. Different approaches to handle imbalanced data-sets include random under-sampling and random over-sampling techniques [146]. Random under-sampling aims to balance class distribution by randomly eliminating majority class samples (normal cases). This is done until the majority and minority class instances are balanced out as done in [95]. On the other side, over-sampling increases the number of instances in the minority class (abnormal cases) by randomly replicating them to present a higher representation of the minority class. Unlike under-sampling, over-sampling leads to no information loss.

The appropriate approach (random under-sampling or random over-sampling) depends on the amount of available data-set and the specific problem at hand. Researchers empirically test each approach and select the one that gives them the best results. In the case of using an imbalance

data-set, accuracy is not the right metric to evaluate the performance of the model. There are more appropriate scores when using imbalanced data-sets such as F1-score [46] that combines the trade-offs of precision and recall, and outputs a single number reflecting the goodness of a classifier.

4.3.8 Multi-stage and End-to-End (E2E) Methods

A multi-stage pipeline used for detection and classification of a lesion consists of multiple stages such as pre-processing, image segmentation, feature detection, feature selection, and classification stages [63,177]. End-to-end (E2E) deep learning methods take all these multiple stages and replace it with just a single neural network. Researchers in [33, 44, 59, 75, 122, 140, 188, 211, 214, 230] have used one or more stages of this multi-stage pipeline in their CNN systems. In their multi-stage method, CNN is trained to determine whether a small patch has Mass and/or MCs. Other researchers focused on training a deep CNN for classifying a small ROI or full image into benign or malignant, assuming an existing Mass/MCs detection system as in [26, 32, 58, 62, 71, 106, 108, 111, 112, 194, 218, 244]. In multi-stage methods for CNNs, several cascaded classifiers are trained independently, each classifier makes a prediction, and all predictions are combined into one using different strategies. Dhungel et al. have found that the multi-stage methods are effective in the reduction of false-positive detection [56]. Moreover, researchers in [17, 57, 58, 79, 124, 174, 238, 258] used the E2E methods.

E2E methods for MGs are better than multi-stage methods when training a CNN with a large data-set. But if the data-set is small in size, then the learning algorithm cannot capture much insight from data. Excluding potentially useful hand-crafted features that are very helpful if well designed is the downside of the E2E approaches. Therefore, the key parameter to choose using the E2E deep

learning approach is having sufficient data to learn the model.

4.4 Toolkits and Libraries for Deep Learning

Implementing a DL network from scratch is an exhausting process and probably beyond the skills of most medical imaging researchers. It is much more efficient to utilize the publicity available resources. Some criteria should be considered while choosing a library and toolkit including its programming language for the interface, the quality of documentation of the toolkit, the ease of programming, the runtime to do thousands of calculations per pixel, the training speed, GPU support for faster performance [84], and lastly its popularity among experts. Recent surveys were done in [64, 197] discusses the most famous and recent toolkits and libraries used generally for DL. The common toolkits used in training CNNs for mammography are Tensorflow [2], Keras, Caffe [109, 110], PyTorch [166] and MatConvNet [225]. Table 4.1 gives a comparison between these libraries and their ranking based on the forks received by the community on GitHub.

4.4.1 Tensorflow

Tensorflow is one of the most popular DL libraries, it was developed by the Google Brain team and open-sourced in 2015 [2]. Tensorflow is a Python-based library capable of running on multiple CPUs and GPUs. It can be used directly to create deep learning models, or by using wrapper libraries (e.g. Keras) on top of it. Tensorflow does not contain many pre-trained models and there's no support for external data-sets, like Caffe. The framework is written in C++ and Python and has a large amount of available documentation. As of today, it is the most commonly used deep learning framework.

4.4.2 Keras

Keras is a very lightweight open-source library, easy to use, and pretty straightforward to learn. It was built as a simplified interface for building efficient deep neural networks in just a few lines of code and use Tensorflow as back-end.

4.4.3 Caffe

Caffe is one of the first deep learning libraries developed mainly by Berkeley vision and learning center (BVLC) [109, 110]. It is a C++ library that also has a Python interface and finds its primary application in modeling CNNs. Caffe provides several pre-trained networks directly from the Caffe Model Zoo, available for immediate use.

4.4.4 PyTorch

PyTorch is a Python library enabling GPU accelerated tensor computation, similar to NumPy. A few advantages of using PyTorch are its multi-GPU support, dynamic computational graphs, custom data loaders, optimization of tasks, and memory management. PyTorch provides a rich API for neural network applications [166]. PyTorch is used by many companies such as Twitter, Facebook, and Nvidia to train DL models.

4.4.5 MATLAB

MATLAB has a neural network toolbox that provides algorithms to create, train, visualize deep neural networks. TL can be done with pre-trained deep CNNs models (including Inception-v3, ResNet-50, ResNet-101, GoogLeNet, Alex-Net, VGG-16, and VGG-19) and models imported

from Keras or Caffe. MATLAB allows computations and data distribution across multi-core processors and GPUs with the parallel computing toolbox. MatConvNet [225] is an open-source implementation of CNNs with deep integration in the MATLAB environment.

Table 4.1: Comparison between most famous toolkits and libraries for training mammography.

	Interface	**Languages**	**Open source**	**CUDA support**	**Pre-trained models**	**Forks (Github)**	**Contributions (Github)**
TensorFlow	Python	C++, Python	Yes	Yes	Yes	63,603	1,481
Keras	Python, R	Python	Yes	Yes	Yes	11,203	681
Cafee	Python, Matlab, C++	C++, Python	Yes	Yes	Yes	14,868	267
PyTorch	Python	C, Python, CUDA	Yes	Yes	Yes	3,592	644
MatConvNet	Matlab	CUDA	Yes	Yes	Yes	651	24

Chapter 5

Applications of Deep CNNs for Mammography

5.1 Applications

After describing deep CNNs in the previous chapters, and different practices that are famous for mammography, we will now turn our focus to how these are used for recognition purposes for mammography. More specifically, we review recent deep CNNs' applications in mammography such as classification, localization, image retrieval, high-resolution image reconstruction, and risk analysis. We summarized these recent works in Table 6.2. In the Classification task, the model is given an image and k possible classes. The task is to decide which of the k classes the image belongs to. For example, a medical image contains either no tumor, benign tumor or malignant tumor: Which of those given three classes is the image?. In the localization task, the algorithm is given an image and one class k. The task is to find bounding boxes for all instances of k. For example, if the medical image has multiple tumors, the localization model shown find bounding boxes surround these tumors. In contrast, the detection task is given an image and k classes, and the output is to find bounding boxes for all instances of those classes. In the semantic segmentation, the model is given an image and k classes, and it is required to classify each pixel independently. But in instance segmentation, the model is given an image and k classes, classify each pixel as

one of the k classes, but distinguish different instances of the classes. For example, if the image contains several cars, the model will assign a color to every car in the image. In the content-based Image Retrieval, the model is given an image x and n images in a database, and it is required to find the top u images that are most similar to x. For example, the model is given an image that contains a malignant tumor and it is required to find images that have similar features as the given one.

5.1.1 Lesion Classifications and Detection

The detection of lesions in mammography is a common task for CNNs. In contrast to lesion detection, classification of MGs into benign and malignant is a challenging task that many studies try to address it. The authors in [33, 36, 44, 55, 59, 76, 121, 122, 140, 154, 188, 211, 213, 214, 230, 236, 256] are interested in lesion classification into two classes. They developed a CNN to predict a probability of being normal (NL), contain mass and/or MCs. The studies in [3, 12, 24–26, 28, 32, 48, 62, 71, 89, 103, 106–108, 111, 112, 125, 135, 175, 176, 190, 192, 195, 218, 244] present deep CNN methods to classify the MG images into 2 classes (benign or malign), or three classes (benign, malign or without tumor). The authors in [40, 41] studied the development of malignancy of mass(es). The authors in [44, 75, 154, 223, 230] are interested in the classification and detection of MCs in mammography. Chan et al. [44] introduced one of the earliest applications of CNNs to detect clustered MCs. The authors applied enhancement filters for noise reduction on fifty-two FSM images. They observed that the shape of MCs in the breast is randomly oriented, thus they introduced an augmentation technique. Sahiner et al. [188] demonstrated the great effect of mixing CNN representation features and textural features (AUC of 0.873). Lo et al. [140] introduced a multiple circular paths

CNN coupled with morphological features of ROIs (AUC of 0.89). Sharma et al. [195] extracted geometrical features from MG images and used it with the representation features of their CNN. Their work demonstrates that DL methods are superior to traditional classifiers. Domingues et al. [59] used a shallow CNN that did not outperform traditional CAD methods, as they used a very small data-set to train their network and the selected normal ROIs did not represent every possible aspect of healthy breast tissue. Antropova et al. [24] developed a system incorporating both deep CNN and conventional CAD methods that performed statistically better than either one separately.

Sert et al. [192] stated that human-level recall performance in detecting breast cancer considering MCs from MGs has a recall value between 74.5% and 92.3%. In [192], the authors reached a recall value of 94.0% above human-level performance. Wang et al. [230] showed that breast arterial calcifications (BACs), detected in MGs, can be useful for identifying risk markers for having cancer. The authors in [230] showed that their CNN method achieves a level of detection similar to the human experts. Kooi et al. [122] employed a deep CNN with a large augmented data-set. Similar to the work of [230], the network in [122] performs similar to experienced radiologists, achieving AUC of 0.87 while the mean AUC of the experienced radiologists is 0.84. In [124], Kooi et al. proposed to use a random forest classifier for mass detection followed by a deep CNN that classifies each detected mass. Their method relies on manually extracted features and features extracted from CNN layers. In [124], Kooi et al. trained their model on a large data-set and integrated additional information such as lesion location and patient information. Kooi et al. [125] following their work in [122,124] employed a conditional random field (CRF) that is trained on top of CNN to model contextual interactions such as the presence of other suspicious regions. In [121], Kooi et al. employed a deep MV CNN using a pre-trained network on a medical domain. They

combined the extracted features using the deep CNN with hand-crafted features.

The studies in [71, 87, 103, 111, 113, 135, 189, 190, 213] demonstrated the use of TL in their work. The authors in [71, 103, 111] showed that CNNs in addition to TL can superior current CAD methods for tumor detection and classification based on small data-sets. Samala et al. [189] demonstrated that MGs can be useful for pre-training a deep CNN for mass detection in digital breast tomosynthesis (DBT). The similarity between masses in mammography and DBT can be observed from the ability of the DCNN in recognizing masses in DBT. In [190], Samala et al. demonstrated that CNNs with TL achieve better generalization to unknown cases than networks without TL. Similar to [189,190], Hadad et al. [87] described a TL approach for using a pre-trained deep CNN on MGs to improve the detection accuracy of fine-tuned CNN on breast MRI lesions. Suzuki et al. [213] developed a deep CNN pre-trained on natural images, then the authors modified the last fully connected layer and subsequently train the modified CNN using 1,656 ROIs. Similar to [213], Jiao et al. [112] achieved an accuracy of 96.7% by applying fine-tuning on a pre-trained CNN on natural images to extract features for the next procedures. Jiao et al. [113] following his work in [112] proposed metric learning layers to further improve the performance of the deep structure and distinguish malignant instances from benign ones. Levy and Jain [135] demonstrated that a fine-tuned pre-trained network significantly outperforms shallow CNNs.

Abbas [3] used speed-up robust features and local binary pattern variance descriptors that are extracted from ROIs. After that, they constructed deep invariant features in supervised and unsupervised fashions through a multilayer CNN architecture. Valvano et al. [223] achieved accuracy of 83.7% for MCs detection using a deep CNN. Jamieson et al. [108] introduced a four-layer unsupervised adaptive deconvolution network to learn the image representation using 739 FFDM

images. Sun et al. [211] developed graph-based semi-supervised learning (SSL) method using a deep CNN, their method allows the users to include the unlabeled data into the DL training dataset. In contrast, Arevalo et al. [25] used supervised training in their method using ROIs annotated manually made by expert radiologists, achieving AUC of 0.86. Arevalo et al. [26] following their work in [25], used a hybrid supervised CNN classifier along with an extensive enhancement pre-processing process.

Dubrovina et al. [62] presented a supervised CNN for region classification into semantically coherent tissues. The authors overcame the difficulty involved in a medium-size database by training CNN in an overlapping patch-wise manner. Teare et al. [218] proposed dual supervised CNNs for classifying full MG images to normal, benign and malignant classes. In their work a random forest classifier was trained, taking the outputs of the two-deep CNNs.

The authors in [12, 25, 36, 48, 75, 106, 107, 111, 154, 236, 258] applied pre-processing, augmentation, normalization, regularization, mixing FSM and FFDM MG images, and other techniques to better implement their network. Ge et al. [75] compared the performance of CNNs on pairs of FFDM and SFM obtained from the same patients with a period less than 3 months. Their results show that CNN with FFDM images (AUC of 0.96) detect more MCs than the CNN with FSM images (AUC of 0.91). Hepsaug [95] achieved an accuracy of 88% when training separate deep CNN on only mass ROIs and 0.84% on training deep CNN on only MCs ROIs in the BCDR database. On the other hand, the accuracy results show that classifying only mass or only MCs is more successful compared to classifying mass and MCs data. Zhu et al. [256] conducted mass detection for whole MG images. Their deep multi-instance network uses linear regression with weight sharing for the malignant probability of each position from CNN's feature maps. The authors in [12, 71]

trained a multi-stage CNN network for the classifications of lesions in MGs. Bekker et al. [32] presented a deep MV CNN for the classification of clustered breast MCs into two classes. Their results show that classification based on MV MGs shows promising results. Carneiro et al. [40] addressed the classification of mass(es) using a pre-trained MV CNN. Their model classifies a full MG by extracting features from each view of the breast (train a separate CNN for each view) and combining these features in a joint CNN model to output a prediction that estimates the patient's risk of developing breast cancer. Carneiro et al [41] following his work [40] build a fully automated pre-trained CNN for detecting masses and MCs in MV MG images. Geras et al. [79] developed an MV CNN that utilizes large high-resolution images without downscaling. They showed that the accuracy of detecting and classifying MGs clearly increases with the size of the training data-set and that the best performance can only be achieved using the images in the original resolution. Yi et al. [244] utilized a deep MV learning by averaging the probability scores of both views to make the final prediction. Lotter et al. [143] introduced a multi-scale deep CNN trained with a curriculum learning strategy. Lotter et al. first train CNN-based patch classifiers on ROIs, and then use the learned features to initialize a scanning-based model that renders a decision on the whole image, having final results by averaging final scores across MV of the breast. Dhungel et al. [56,57] presented a cascade DL network for detecting, segmenting and classifying breast masses from MGs with minimal user intervention. Dhungel et al. [58], following their work in [55–57], implemented an MV deep residual neural network for the fully automated classification of MGs as either malignant or normal/benign (AUC of 0.8).

5.1.2 Risk Assessment

The studies in [15,31,70,115,153,220,238] have demonstrated that applying CNNs methods have significant potential to develop a new short-term risk predicting scheme with improved performance in detecting early abnormal symptom from the negative MGs. Breast density is considered a strong indicator of breast cancer risk [15,31,69,70,115]. Fonseca et al. [69,70] explored an automatic breast composition classification work-flow based on CNN for feature extraction in combination with a support vector machines classifier. A similar approach was done by Becker [31] achieving an (AUC of 0.82) comparable to experienced radiologists (AUC of 0.79–0.87).

Li [137] trained a deep CNN to estimate a probability map of breast density (PMD) to classify mammographic pixels into a fatty class or dense class. Kallenberg et al. [115] presented an unsupervised CNN for breast density segmentation and automatic texture scoring. The model learns features across multiple scales, then they are fed to a simple classifier that is specific to the task of interest yielding AUC of 0.59. Ahn et al. [15] used CNN for the task of automatic classification of mammographic breast tissues into dense and fatty tissues. Their CNN is configured to learn the local features from image patches while keeping the context information of the whole MG. Wu et al. [238] managed to train an MV deep CNN using a data-set of 201,179 MGs for breast density classification. Mohamed et al. [153] achieved AUC of 0.95 when using only the MLO view images. In comparison, the AUC is 0.88 when using only the CC view images. When both the MLO and CC view images were combined as a single data-set, the AUC is lowered to 0.92. The authors in [151] following their work in [153] achieved a better AUC of 0.98 by fine-tuning a pre-trained network. Hang [89] achieved a classification accuracy of 66% for classification of full images into normal, benign and malignant.

5.1.3 Lesion Localization

For localization, the information about which category an image belongs to is already available and the task is to instead figure out where exactly the object is located in the image. Classification and localization can also be combined so that a fixed amount of lesions in an image will be classified and also located. This task called multi-class localization. The following authors employed CNNs in the aim of lesions classification and then localization within these images [17, 21, 43, 49, 55, 57, 65, 119, 169, 212, 258], potentially enabling E2E training. Ben-Ari et al. [33] introduced the detection of AD using a supervised pre-trained region-based network (R-CNN). Ertosun and Rubin [65] developed an E2E dual CNN based visual search system for localization of mass(es) in MGs. Kisilev et al. [119] gave a semantic description for MGs. The authors presented a multi-task R-CNN approach for detection and semantic description of lesions in diagnostic images. Carneiro and Bradley [43] presented an automated supervised architecture composes of a multi-scale deep belief network that selects suspicious regions to be further processed by a two-level cascaded R-CNN. Akselrod et al. [17] integrated several cascaded segmentation modules into a modified cascaded R-CNN. Hwang et al. [104] proposed a self-transfer learning framework that enables training CNNs for object localization without neither any location information nor pre-trained models. Zhu et al. [258] introduced an E2E adversarial training for mammographic mass segmentation to learn robustly from scarce MGs. The authors highlighted the importance of pre-processing, augmentation, image enhancement, and normalization techniques. The authors stated that it is not feasible to use networks pre-trained on general images since ROI characteristics of medical images are thoroughly different from natural images. However, their opinion contradicts other researchers work.

The authors in [49,55,212] proposed a patch-based CNN to detect masses. Choukroun et al. [49] proposed a method that classifies MGs by detecting discriminative local information contained in patches through a deep CNN and then uses the local information to localize tumors. Dhungel et al. [55] used the output from a CNN as a complimentary potential function to a deep belief network (DBN) models for the localization of breast masses from MGs, using a small training data-set. A drawback of the patch-based approach in [55, 212] is that the input patches came from non-overlapping areas, which makes it difficult to preciously localize masses. Moreover, the size of the input patches in [55, 212] is very small that produces difficulty in differentiating normal tissues from abnormal ones.

The authors in [21, 169] used the famous YOLO-based deep CNN [181] for breast mass classification and localization. The trained YOLO-based system localizes the masses and classifies their types into benign or malignant. The authors in [21] achieved a mass location with an overall accuracy of 96.33% and detection of benign and malignant lesions with an overall accuracy of 85.52%.

5.1.4 Image Retrieval

Tasks like medical image retrieval using DL have been lately addressed in the medical field to facilitate the process of production and management of large medical image databases. Conventional methods for analyzing medical images have achieved limited success, as they are not capable to tackle the huge databases. The learned features and the classification results from training CNN are used to retrieve medical images. Qayyum et al. [173] proposed a DL based framework for content-based medical image retrieval (CBMIR) by training a deep CNN for the classification tasks using

medical images for different body organs (e.g. MGs, lungs, brain, liver, etc. Qayyum et al. [173] achieved an average classification accuracy of 99.77% for 24 classes of medical images. Similarly, Ahmad at el. [14] trained a deep CNN for CBMIR of different 193 classes for different body organs. Moreover, [14] applied TL and augmentation to increase the performance of their deep CNN.

5.1.5 Super Resolution Image Reconstruction

The task of super-resolution image reconstruction using CNN (SRCNN) is an E2E mapping between the low and high-resolution images for enhancing images [61]. The mapping is represented as a deep CNN that takes the low-resolution image as the input and outputs the high-resolution one. The study of Umehara et al. [222] shows that SRCNN can significantly outperform conventional interpolation methods for enhancing image resolution in digital mammography especially in dense breasts.

Chapter 6

Research Challenges and Directions for Mammography

6.1 Abstract

The limitations of traditional computer-aided detection (CAD) systems for mammography, the extreme importance of early detection of breast cancer and the high impact of the false diagnosis of patients drive researchers to investigate deep learning (DL) methods for mammograms (MGs). Recent breakthroughs in DL, in particular, convolutional neural networks (CNNs) have achieved remarkable advances in the medical fields. Specifically, CNNs are used in mammography for lesion localization and detection, risk assessment, image retrieval, and classification tasks. CNNs also help radiologists providing more accurate diagnosis by delivering precise quantitative analysis of suspicious lesions.

In this research, we conduct a detailed review of the strengths, limitations, and performance of the most recent CNNs applications in analyzing MG images. It summarizes 83 research studies for applying CNNs on various tasks in mammography. It focuses on finding the best practices used in these research studies to improve diagnosis accuracy. This survey also provides a deep insight into the architecture of CNNs used for various tasks. Furthermore, it describes the most common publicly available MG repositories and highlights their main features and strengths. The

mammography research community can utilize this survey as a basis for their current and future studies. The given comparison among common publicly available MG repositories guides the community to select the most appropriate database for their application(s). Moreover, this survey lists the best practices that improve the performance of CNNs including the pre-processing of images and the use of multi-view images. Also, other listed techniques like transfer learning (TL), data augmentation, batch normalization, and dropout are appealing solutions to reduce overfitting and increase the generalization of the CNN models. Finally, this survey identifies the research challenges and directions that require further investigations by the community.

6.2 Background

Breast cancer is the second most common cause of cancer death in women. According to the American cancer society's latest statistics, it is estimated that 40,610 women in the USA are expected to die in 2017 from breast cancer. As of March 2017, there are more than 3.1 million women with a history of breast cancer in the USA [198]. Mammography is one of the most widely used methods for breast cancer screening and has contributed significantly to the reduction of the mortality rate through early detection of cancer [138]. However, the complexity of mammograms (MGs) and the high volume of exams per radiologist can result in false diagnoses [66, 235].

Computer-aided detection (CAD), which employs image processing techniques and pattern recognition theory, has been introduced to provide an objective view to radiologists [138]. Studies have shown the effectiveness of CAD models; however, accurate detection of breast cancer has remained challenging [138]. Recent studies show that CAD models cannot improve significantly the diagnostic accuracy of mammography [133]. The biggest challenge in using CAD for abnor-

mality detection in MGs is the high false-positive rates (FPR). False positives result in patient anxiety, additional radiation exposure, unnecessary biopsies, high callback rates, increased health care costs, and additional assessment [235]. In the USA, millions of women undergo screening mammography each year, as a result, even a small reduction in the FPR result in a widespread benefit [90, 198]. The limitations of current CAD indicate the need for new, more precise detection methods.

Recent advances in computational technologies, significant progress in machine learning and image processing techniques, and prevalence of digital MG images have opened up an opportunity to address the challenging issue of early detection of breast cancer using deep learning (DL) methods [129, 131, 139, 252]. Recently, DL methods, especially convolutional neural networks (CNNs, also known as ConvNets) have gained lots of attention to CAD for MGs as they help overcome CAD systems' limitations [94, 138, 139, 252]. CNNs achieve higher detection accuracy than CAD models, and help radiologists provide a more accurate diagnosis by delivering quantitative analysis of suspicious lesions [45, 122, 131, 169]. A recent research study shows that using DL methods drop the human error rate for breast cancer diagnoses by 85% [230]. Current CNN models are designed to improve radiologists' ability to find even the smallest breast cancers at their earliest stages alerting the radiologist to the need for further analysis [122, 230].

Recent studies used CNNs to generate a standard description of lesions, which can help the radiologist in making a more accurate decision [122, 169]. Moreover, advances in CNNs can not only aid radiologists but also eventually make diagnosis systems to read MGs independently soon [122]. In the last few years, CNNs have led to breakthroughs in a variety of pattern recognition and classification problems for natural images due to the availability of big data repositories, fast

graphical processing units, and the power of parallel and distributed computing [84, 127, 129, 131].

Training a deep CNN model with a limited number of medical data is very challenging, which has been addressed by using transfer learning (TL) and augmentation techniques [127, 129, 234]. Studies show that CNN methods that compare images from left and right breasts [50] and also the craniocaudal (CC) and mediolateral-oblique (MLO) view of each breast can improve the accuracy of detection and reduce the false positives [33, 58, 79, 121, 230, 244, 256]. CNNs have also been used in the risk assessment applications to increase the accuracy of early detection breast cancer by radiologist [15, 31, 40, 70, 115, 136, 163, 168, 174, 212]. In this work, we summarize almost all contributions, as of November 2017, to the field of DL in MGs, in particular using CNNs.

6.3 Material and Methods

6.3.1 Criteria for Inclusion/Exclusion of Studies in the Survey

We carried out comprehensive literature research, using the defined keywords given in Table 6.1, on journals and proceedings of scientific conferences including, but are not limited to the following scientific databases: Scopus, ACM Digital Library, Science Direct, IEEE Explore Digital Library, PubMed, Web of Science. In total, we considered 83 studies from the period of 1995 to Nov 2017. These studies focus on implementing CNNs for lesion localization and detection, risk assessment, image retrieval, high-resolution image reconstruction and classification tasks in MG images. The inclusion/exclusion criteria we used for this review are presented in Table 6.1. Figure. 6.1, shows a breakdown of the studies included in this survey in the year of publication grouped by their neural network task.

In this study, we addressed the following research questions:

Table 6.1: Inclusion/exclusion for the systematic review.

Category	Criteria
Time period	Published from 1995 to the present (Nov 2017).
Databases	Private and public databases.
Publication	English articles in print. Excluded articles accepted for publication before appearance in journals or conferences as of Aug 2017.
Research focus	All implementation of CNNs for breast cancer in Mammography.
Keywords	Deep learning, convolutional neural networks, breast cancer, mammography and transfer learning.
Abnormalities	Mass, calcification, architectural distortion and asymmetries.

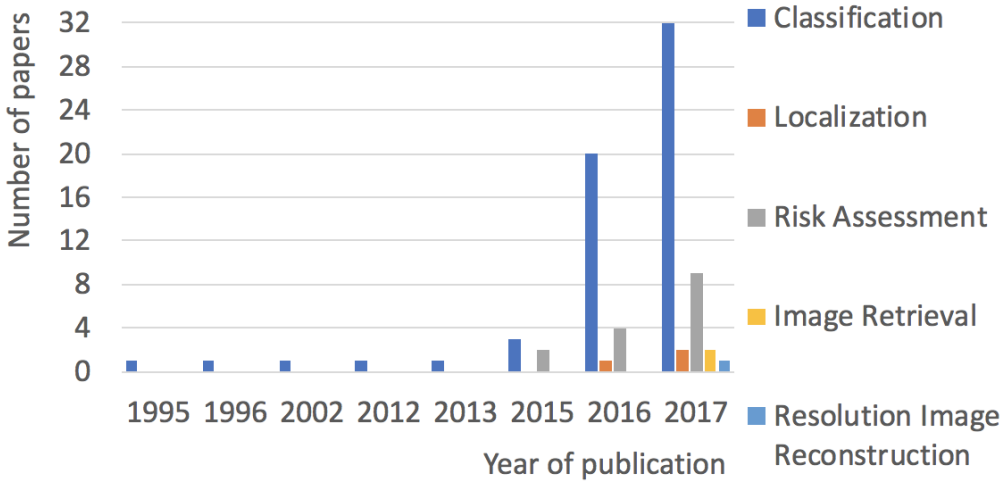


Fig. 6.1: A breakdown of the studies included in this survey in the year of publication grouped by their neural network task. Since 2016 the number of studies on CNN for MGs has increased significantly.

- Does this study focus on using a CNN for detecting abnormalities in MGs?
- What is the task of the implemented CNN?
- What are the databases, database size, image resolution, image type, abnormalities involved in the development of the CNN?

- What are the methodologies used for the setup and pre-processing of the data-set?
- Can deep networks perform well on medical images specifically MGs?
- What are the learning methods used for training the CNNs?
- What are the best practices that were applied to increase the accuracy of the detection of abnormalities?
- What are the advantages and limitations presented by the methodologies employed in CNNs?
- Is it an end-to-end (E2E) training method?
- Is transfer learning from natural imagery to the medical domain-relevant?
- Is combining learned features with hand-crafted features will enhance the accuracy of a certain mammographic task?
- What are the common toolkits used in mammography?
- What are the challenges to train a deep neural network for mammography data-set?
- How imbalanced data-sets impact the performance of CNNs?
- What is the common cross-validation method used with MGs?
- Which activation functions are commonly used for training MGs?

6.4 Research Challenges and Directions

In this section, we list the research challenges and directions that require further investigations by the community.

6.4.1 Localization of Tumors

The patch-based CNNs, R-CNNs, Fast R-CNNs, Faster R-CNNs, and YOLO methods have recently become more popular for localization tasks for MGs. Faster R-CNN is the choice of most of the mammography researchers who aim to obtain high detection accuracy numbers. However, training R-CNN and its variants faster versions are time-consuming and memory expensive. In contrast, for faster computations, less accurate detection, and limited memory computations, the YOLO method is the right choice. Finally, patch-based CNN methods are not recommended and result in many false positives. More research needs to be done for better localization of tumors in MGs.

6.4.2 Limited Data for Learning

One of the challenging problems that face researchers while training CNNs is the size of the training data-set. As discussed in the best practice section, although several approaches such as data augmentation, TL, and drop out have been used to handle the problem of training the model with limited samples, this problem has remained challenging.

6.4.3 Imbalanced Data-set

Another challenging problem is the imbalance ratio between positive and negative classes in the training data-sets. Training CNN models directly on imbalanced data-sets may bias the prediction towards the more common classes like normal. The effect of imbalanced data-set on the performance of a CNN for MGs has not been studied thoroughly. Some works used a balanced data-set and some used imbalanced ones. Since in general less abnormal MGs are available to compare

to normal MGs it is very important to investigate the effect of using balanced and imbalanced data-sets on the accuracy of the CNN model.

6.4.4 Size of Lesions

The size variation of lesions within MG images is another challenge for training CNNs in detecting cancer. Resizing a large MG to 224×224 or 227×227 (common choices among researchers) will likely make the ROI hard to detect and/or classify. To address this problem, several studies have proposed to train a CNN model using different scales of lesions [43, 115, 143]. More research is required to find lesions of different sizes.

6.4.5 Memory Constraints

The classification of the whole size MG images is challenging due to the memory constraints and increased feature space. Researchers in [79, 251] address this problem by resizing the images to smaller ones, however, this affects the accuracy of their model. More research should be done on how to overcome the memory constraints while training CNNs with full-size MG images.

6.4.6 Non-annotated Data-set

Another challenging problem to researchers is how to train a CNN model using a non-annotated data-set. In a non-annotated data-set, the input image to the CNN model is binary labeled as normal or cancerous without any details about the location of the abnormalities. To address this problem, Lotter at el. [143] train a patch-level CNN classifier, which is then used as a feature extractor to an image-level model. Training the CNNs for classification of non-annotated data-set is still an open area for research [143, 196, 256].

6.4.7 False Positives Reduction

Even though CNNs are very successful in providing better performance compared to traditional CADs, they still result in false positives. False-positive results cause patients needless anxiety, additional testing, biopsies, and unnecessary costs. Several approaches have been proposed to improve false positive in CNNs such as using MV CNNs [33,41,58,79,121,123,230,238,244,256]. However, more research is required to integrate prior images with current screening to eliminate false positives.

6.4.8 Multiple Detection

Current CNN models are trained to detect and/or localize mass(es) within MGs neglecting the existence of MCs. More research should be directed on detecting multiple abnormalities within the same breast.

6.4.9 Pre-processing Filters

In FSM images, a significant number of abnormalities are misdiagnosed or missed due to the less visibility, low contrast, poor quality, and noisy nature of these images. Common pre-processing techniques (e.g. CLAHE, median filter) are proposed in [5,36,117,218] to enhance image quality, image smoothing and noise reduction. However, choosing the proper pre-processing technique for MGs to improve the classification of CNNs is still an open problem.

6.5 Discussion and Recommendations

We show a breakdown of the studies included in this survey grouped by their neural network task (see Table 6.2). Figure. 6.2, shows the percentage of studies employing some of the CNN best practices that are discussed in the previous section and are shown in Table 6.2. 78 studies (out of 83) used common pre-processing techniques to enhance the quality of images, reduce or remove noise, and improve the contrast of MGs. That shows the importance of having a good separation between foreground and background pixels and not removing the important information from the images. Moreover, 59 studies used ROIs for more efficient computation, while 23 studies applied CNN to MG of full image size as in [17, 40, 70, 79, 104, 218, 256, 258]. Even for CNNs that are trained with full image size, the pre-processing is mandatory to remove marks, labels, pectoral muscle and black areas that can interfere in the post-processing of these images. Data augmentation has been recommended and employed by 52 studies. Data augmentation reduces overfitting by generating more instances of training data. TL is gaining more popularity for medical images, 32 studies have successfully applied it to pre-train their network. From 2015 until now, there is an increasing trend in using TL. 15 studies implemented MV CNNs which lead to significant improvements in the performance of the single-view ones. It is a beneficial practice to use both CC and MLO views to detect abnormalities. 25 studies implemented an E2E CNN which may include segmentation, detection, and classification of lesions in MGs. We summarize the recommendations to significantly improve the performance of CNNs in detection and classification of breast cancer using MG images as follows:

- Use pre-processing techniques such as CLAHE filter to improve the contrast of MGs, median filter to reduce noise, and un-sharp masking to smooth the images.

- Apply cropping and down-sampling for more efficient computation.
- Use a suitable validation approach according to the size of the data-set available.
- Use augmentation, drop-out, and TL to reduce overfitting and increase the generalization of the model.
- Use suitable batch size if using ROIs.
- Use multi-view (MV) CNNs to embed more information for better performance.
- Use full resolution images if it is computationally practical.
- Mix between FFDM and FSM images.
- Use suitable activation functions such as ReLU, be careful with initializing the learning rates and possibly monitor the fraction of dead neurons in the network.
- Use large well-labeled data-set if available.
- Go deeper in layers if a large data-set is available.
- Use context and patient information in multi-modal models.
- Use recently available libraries for implementing CNNs such as Tensorflow or Keras.

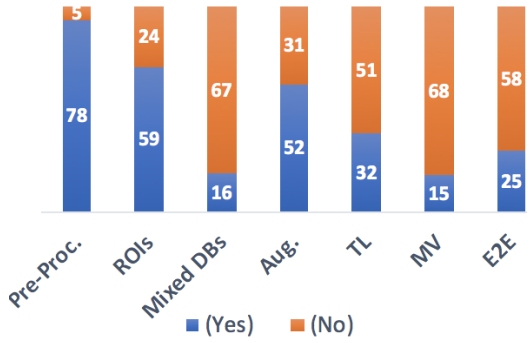


Fig. 6.2: Statistics for the included studies.

Table 6.2: A comparison between different approaches in literature from the point of view 1) author, 2) database used, 3) number of images used before and after augmentation in parenthesis, 4) the use of full image or ROI, 5) if pre-processing is done, 6) size of images used in pixel, 7) if augmentation is applied, 8) if it is end-to-end (E2E) technique, 9) if transfer learning (TL) is used, 10) muli-view (MV) or single-view (SV), 11) area under curve (AUC), 12) accuracy (Acc.%), 13) class used: normal images (NL); benign images (B); malignant (M); calcifications (MCs); architectural distortion (AD), 14) lesion being segmented, 15) number of layers in the architecture.

Author	Year	Database	#Images	Kind	Pre-pro.	Size	AUG.	E2E	TL	View	AUC	Acc.%	Class	Lesion	Layers
Lesion classification															
Sahiner [188]	1996	Private	168(672)	ROIs	Y	32×32	Y	N	N	SV	0.87	-	NL-Mass	Mass	3
Lo [140]	2002	MIAS	200(3,193)	ROIs	Y	512×512	Y	N	N	SV	0.89	-	NL-Mass	Mass	3
Domingues [59]	2013	INbreast	112	ROIs	N	32×32	N	N	N	SV	0.89	-	NL-Mass	Mass	3
Dhungel [55]	2015	BCRP, INbreast	116-77	ROIs	N	40×40	N	N	N	SV	0.9	-	NL-Mass	Mass	6
Kooi [122]	2016	Private	398(2.5M)	ROIs	Y	250×250	Y	N	N	SV	0.87		NL-Mass	Mass	11
Wichakam [236]	2016	INbreast	216(1,728)	ROIs	Y	32×32	Y	N	N	SV	-	98.4	NL-Mass	Mass	9
Zhu [256]	2017	INbreast	410	Image	Y	224×224	Y	Y	Y	MV	0.85	90	NL-Mass	Mass	8
Suzuki [214]	2016	DDSM	1,656	ROIs	Y	454×454	N	N	Y	SV	-	-	NL-Mass	Mass	8
Dubrovina [62]	2016	N/A	40(8k)	ROIs	Y	256×256	Y	N	N	SV	-	-	NL-Mass	Mass	7
Suzuki [213]	2016	DDSM	1,656	ROIs	Y	454×454	N	N	Y	SV	-	-	NL-Mass	Mass	8
Kooi [121]	2017	Private	398(2.5M)	ROIs	Y	250×250	Y	N	Y	MV	0.87	-	NL-Mass	Mass	22
Sun [211]	2017	Private	3158(25K)	ROIs	Y	52×52	Y	N	N	SV	0.88	82.2	NL-Mass	Mass	7
Samala [189]	2016	Private	2461(45072)	ROI	Y	128×128	Y	N	N	SV	0.99	-	NL-Mass	Mass	8
Hadad [87]	2017	Private	11092(32,064)	ROI	Y	32×32	Y	N	N	SV	0.94	87	NL-Mass	Mass	13
Chan [44]	1995	Private	52(1,744)	ROIs	Y	16×16	Y	N	N	SV	0.91	-	NL-MCs	MCs	3
Gardezi [74]	2017	IRMA	2795(25k)	ROIs, Image	Y	224×224	N	N	N	SV	1	100	NL-Mass	Mass	16

Continuation of Table 6.2

Author	Year	Database	#Images	Kind	Pre-pro.	Size	AUG.	E2E	TL	View	AUC	Acc.%	Class	Lesion	Layers
Lesion classifications															
Ge [75]	2007	Private	192	ROIs	Y	16×16	Y	N	N	SV	0.96	-	NL-MCs	MCs	3
Valvano [223]	2017	Private	290(90,000)	ROIs	Y	99×99	Y	N	N	SV	-	83.7	NL-MCs	MCs	10
Mordang [154]	2016	Private	1606(45M)	ROIs	Y	13×13	Y	N	N	MV	-	-	NL-MCs	MCs	9
Wang [230]	2017	Private	840(1.94M)	ROIs	Y	95×95	Y	N	N	MV	-	-	NL-MCs	MCs	12
Bria [36]	2017	Private	7,579(≈27M)	ROIs	Y	12×12	Y	N	N	SV	-	-	NL-MCs	MCs	8
Ben-Ari [33]	2017	DDSM	136(21k)	ROIs	Y	224×224	Y	N	Y	MV	-	80	NL-AD	AD	9
Jamieson [108]	2012	Private	739(2,393)	ROIs	Y	140×140	Y	N	N	SV	0.71	-	B-M	Mass	4
Arevalo [25]	2015	INbreast	736(5,152)	ROIs	Y	150×150	Y	Y	Y	SV	0.82	-	B-M	Mass	7
Agarwal [12]	2015	DDSM	8752(50k)	ROIs	Y	64×64	Y	N	N	SV	-	90	B-M	Mass, MCs	8
Jiao [112]	2016	DDSM	600(1,800)	ROIs	Y	227×227	Y	N	Y	SV	-	96.7	B-M	Mass	9
Bekker [32]	2016	DDSM	1410	ROIs	N	N/A	N	N	N	SV	0.89	78.7	B-M	MCs	3
Sharma [195]	2016	DDSM	40	ROIs	Y	N/A	N	N	N	SV	0.85	79.2	B-M	Mass	N/A
Levy [135]	2016	DDSM	1820(36k)	ROIs	Y	224×224	Y	Y	N	SV	-	92.9	B-M	Mass	9
Abbas [3]	2016	MIAS, DDSM	600	ROIs	Y	250×250	N	Y	N	SV	0.91	91.5	B-M	Mass	N/A
Huynh [103]	2016	Private	607	ROIs	Y	256×256	N	N	Y	SV	0.86	-	B-M	Mass	8
Arevalo [26]	2016	BCDR	736(5k)	ROIs	Y	150×150	Y	N	Y	SV	0.86	-	B-M	Mass	7
Jiang [111]	2017	BCDR	736(N/A)	ROIs	Y	227×227	Y	N	Y	SV	0.88	-	B-M	Mass	22
Samala [190]	2017	DDSM, Private	322(2576) 8334(17056)	ROI	Y	128×128	Y	N	Y	SV	0.82	-	B-M	Mass	10
Sert [192]	2017	DDSM	2620(5965)	ROIs	Y	224×224	Y	N	Y	SV	-	94.3	B-M	MCs	22
Jiao [113]	2017	DDSM	600(1,800)	ROIs	Y	227×227	Y	N	Y	SV	-	92.5	B-M	Mass	8
Jaffar [107]	2017	MIAS, DDSM.	2122(19k)	ROIs	Y	96×96	N	N	N	SV	0.93	93.4	B-M	Mass	8
Chougrad [48]	2017	BCDR	600(N/A)	ROIs	Y	299×299	Y	Y	Y	SV	0.96	97.5	B-M	Mass	221
Bakkouri [28]	2017	DDSM, BCDR.	10k(60k)	ROIs	Y	32×32	Y	N	N	SV	-	97.3	B-M	Mass	7
Antropova [24]	2017	Private	739	ROIs	Y	512×512	N	N	Y	SV	0.86	-	B-M	Mass	19

Continuation of Table 6.2

Author	Year	Database	#Images	Kind	Pre-pro.	Size	AUG.	E2E	TL	View	AUC	Acc.%	Class	Lesion	Layers
Lesion localization															
Qiu [175]	2017	Private	560	ROIs	Y	64×64	N	N	N	SV	0.79	-	B-M	Mass	8
Gallego [71]	2016	MIAS	322(600)	ROIs	Y	227×227	Y	N	Y	SV	-	64.5	NL-B-M	Mass	8
Jadoon [106]	2017	IRMA	2796 (19k)	ROIs	Y	28×28	Y	N	N	SV	-	83.7	NL-B-M	Mass	5
Yi [244]	2017	DDSM	2085(N/A)	ROIs	Y	224×224	Y	N	Y	MV	0.91	85	NL-B-M	Mass	22
Teare [218]	2017	DDSM, ZMDS	>6000(N/A) 1739(N/A)	Image	Y	299×299	Y	N	Y	SV	0.92	-	NL-B-M	Mass, MCs	22
Hepsaug [95]	2017	MIAS, BCDR	1160-2474	ROIs	Y	50×50	N	N	N	SV	-	68, 62	NL-B-M	Mass, MCs	N/A
Hang [89]	2017	DDSM	1318	Image	Y	521×521	N	N	N	MV	-	66	NL-B-M	Mass	13
Kooi [125]	2017	Private	490	ROIs	Y	250×250	N	N	N	SV	0.87	-	Mass-MCs	Mass, MCs	N/A
Qiu [176]	2016	N/A	560	ROIs	Y	512×512	N	Y	N	SV	0.8	-	B-M	Mass	8
Dhungel [56]	2016	INbreast	116(1160)	ROIs	Y	40×40	Y	Y	Y	SV	-	84	B-M	Mass	5
Geras [79]	2017	Private	102,800	Image	Y	$2k \times 2k$	N	Y	N	MV	0.76	-	NL-M	Mass, MCs	11
Kooi [124]	2017	Private	44K(1.3M)	ROIs	Y	250×250	Y	Y	N	SV	0.94	-	NL-M, B-M	Mass	7
Dhungel [58]	2017	INbreast	512	Image	Y	120×120	Y	Y	Y	MV	0.8	-	NL-B-M	Mass, MCs	392
Lotter [143]	2017	DDSM	10480(N/A)	Image	Y	256×256	Y	Y	Y	MV	0.92	-	NL-B-M	Mass, MCs	152
Ribli [184]	2017	DDSM, INbreast	26020(N/A), 115	Image	Y	2.1k×1.7k	Y	Y	N	SV	0.95	-	NL-B-M	Mass, MCs	16
Akselrod [16]	2017	Private	860	Image	Y	224×224	N	Y	Y	SV	0.72	77	NL-B-M	Mass	16
Carneiro [40]	2015	DDSM, INbreast	680-410	Image	Y	264×264	N	Y	Y	MV	0.97	-	B-M	Mass	8
Carneiro [41]	2017	DDSM, INbreast	680-410	Image	Y	264×264	N	N	Y	MV	0.99-0.91	-	NL-B-M	Mass, MCs	8

Continuation of Table 6.2

Author	Year	Database	#Images	Kind	Pre-pro.	Size	AUG.	E2E	TL	View	AUC	Acc.%	Class	Lesion	Layers
Lesion localization															
Kooi [123]	2017	Private	201,851(N/A)	ROIs	Y	250×250	Y	N	N	MV	0.88	-	NL-Mass	Mass	19
Ertosun [65]	2015	DDSM	2420(1.8M)	ROIs	Y	256×256	Y	Y	Y	SV	-	85	NL-Mass	Mass	22
Hwang [104]	2016	MIAS, DDSM	15,837	Image	Y	500×500	N	N	N	SV	0.89	84.1	NL-Mass	Mass	7
Akselrod [17]	2016	Private	850(4,75)	Image	N	800×800	Y	Y	N	SV	-	-	NL-B-M	Mass, MCs	16
Carneiro [43]	2016	BCRP, INbreast	316(3,160), 410(4.1k)	ROIs	Y	40×40	Y	Y	N	SV	-	-	NL-B-M	Mass	8
Choukroun [49]	2017	Private	2,500(28k)	ROIs	Y	224×224	Y	N	Y	SV	0.83	-	NL-B-M	Mass, MCs	8
Zhu [258]	2016	BCRP, INbreast	171(680), 116(464)	ROIs	Y	40×40	Y	Y	N	SV	-	91.3	NL-M	Mass	4
Dhungel [57]	2017	INbreast	410(4.1k)	ROIs	Y	40×40	Y	Y	Y	SV	0.76	91	B-M	Mass	5
Kisilev [119]	2016	DDSM	512	ROIs	Y	128×128	N	N	N	SV	-	-	B-M	Mass	7
Al-masni [21]	2017	DDSM	600	Image	Y	448×448	N	N	Y	SV	0.87	85.5	B-M	Mass	27
Platania [169]	2017	IRMA	10,480(25k)	ROIs, Image	Y	128×128	Y	Y	Y	SV	0.92	93.5	B-M	Mass	19
Sun [212]	2016	Private	420(42k)	ROIs	Y	52×52	N	N	N	MV	0.72	-	B-M	Mass	8
Risk assessment															
Qiu [174]	2016	Private	270	ROIs	Y	256×256	N	N	N	SV	-	71.4	BI-RADS	Density	8
Fonseca [70]	2015	Private	729	Image	N	200×200	N	N	N	SV	-	73	BI-RADS	Density	3
Fonseca [69]	2016	Private	1060(N/A)	Image	Y	200×200	Y	N	Y	SV	-	-	BI-RADS	Density	3
Becker [31]	2017	BCDR	286(N/A)	Image	Y	N/A	Y	Y	N	SV	0.81	-	BI-RADS	Density	N/A
Kallenberg [115]	2016	Private	1,555	ROIs	Y	24×24	N	N	N	SV	-	59	BI-RADS	Density	6
Ahn [15]	2017	Private	10,94(N/A)	ROIs	Y	41×41	Y	N	N	SV	-	-	BI-RADS	Density	16
Li [137]	2017	Private	661(1M)	ROIs	Y	61×61	Y	N	N	SV	-	-	BI-RADS	Density	6
Wu [238]	2017	Private	201,179(N/A)	Image	Y	2.6k×2k	Y	Y	Y	MV	0.93	-	BI-RADS	Density	19
Thomaz [220]	2017	Private	307	Image	Y	260×200	N	N	N	SV	-	98.4	BI-RADS	Density	N/A
Mohamed [153]	2017	Private	15,415	Image	Y	227×227	N	N	N	SV	0.92	-	BI-RADS	Density	8

Continuation of Table 6.2

Author	Year	Database	#Images	Kind	Pre-pro.	Size	AUG.	E2E	TL	View	AUC	Acc.%	Class	Lesion	Layers
Image retrieval															
Mohamed [151]	2017	Private	6000	Image	Y	227×227	N	N	Y	SV	0.98	-	BI-RADS	Density	8
Qayyum [173]	2017	Multiple	7200	Image	Y	224×224	N	Y	N	SV	-	99.8	24 classes	All	8
Ahmad [14]	2017	IRMA	15363(68k)	ROIs	Y	224×224	Y	Y	Y	SV	0.75	-	193 classes	All	16
Resolution image reconstruction															
Umehara [222]	2017	CBIS-DDSM	711	Image	Y	variable	Y	Y	N	SV	-	-	Same image	Any	23

6.6 Summary

We conducted a detailed review of the strengths, limitations, and performance of the most recent CNNs applications in analyzing mammogram (MG) images. This survey systematically compares recent approaches of CNNs in MG images, and show how the advances in DL methods give promising results that can aid radiologists and serve as a second eye for them. The potential role of CNN methods is to handle millions of routine imaging exams, presenting the potential cancers to the radiologists who perform follow-up procedures. We discuss the currently publicly available MG databases. We also give a deep insight into the architectures of CNNs used for various tasks in mammography. This survey represents a valuable resource for the mammography research community since it can be utilized as a basis in their current and future studies. The given comparison among common publicly available MG repositories guides the community to select the most appropriate database for their application(s). Moreover, this survey lists the best practices that improve the performance of CNNs including the pre-processing of images and the use of multi-view

images. Also, other listed techniques like transfer learning (TL), data augmentation, batch normalization, and dropout are appealing solutions to reduce overfitting and increase the generalization of the CNN model. Finally, we identified research challenges and directions that require further investigations for mammography.

Chapter 7

The Effect of Pre-processing on Breast Cancer Detection using CNNs

7.1 Abstract

Recently, advanced CNNs have been developed to conduct different applications in mammography with lower FPR and high detection accuracies. However, due to the low contrast, poor quality, and noisy nature of MG images, a significant number of abnormalities are missed or misdiagnosed. Many pre-processing techniques have been proposed by researchers to enhance image quality, image smoothing and noise reduction. However, choosing a proper pre-processing technique for MGs to improve the classification performance of CNNs is still a difficult task.

7.2 Material and Methods

In this research [5], we have carefully chosen several pre-processing filters for contrast enhancement and noise removal, which are popular in the medical field. We investigated the influence of these pre-processing filters on the classification of MG images using CNNs into three-classes namely normal, benign, and malignant. This can direct the mammography research community to choose the proper pre-processing filter in their future studies. The contrast enhancement filters investigated in our study are CLAHE, fuzzy contrast limited adaptive histogram equalization

(FCLAHE), adaptive unsharp masking (AUM), fast local Laplacian filtering (FLLF). The filters investigated for noise cancellation are adaptive weighted frost filter, adaptive Wiener filter, adaptive median filter (AMF), and adaptive denoising filter.

7.2.1 Proposed Deep learning Model

In this research [5], we adopt the architecture of the popular Alex-net that contains a total of 8 layers, including 5 convolutional layers, and 3 FCLs. We trained two separate networks for each filter independently. The first CNN is trained from scratch, while the second network is pre-trained with about 1.2 million natural images for the classification of 1,000 classes. The last FC layer is then modified to fit our three-classes classification for mammography.

7.2.2 Mammogram Data-sets

We performed experiments on a total of 5,453 ROIs segmented from the publicly available DDSM mammography database. We extracted balanced classes of 2,000 normal patches, 1,954 mass patches (982 benign and 972 malignant) and 1,499 MCs (780 benign and 719 malignant). All the filters are applied one at a time on our augmented data-set for the classification of mass and MCs into three-classes. Moreover, we trained the two CNNs with the same data-set without any enhancement to compare the training results.

7.3 Classification Results

The experimental results show that the training and validation accuracies of the CNN trained from scratch on mass patches increased by 31.52% and 16.87%, respectively, when the FCLAHE filter

is applied; and the FPR dropped from 12.18% to 3.55%. Similarly, for MCs patches, the training and validation accuracies increased by 21.79% and 12.31%, respectively, when the FCLAHE filter is applied; and the FPR dropped from 10.05% to 3.93%. Moreover, the FPR decreased from 12.18% to 3.64% for mass and 10.05% to 8.5% for MCs when adaptive denoising filter and adaptive weighted frost filter are applied to the data-set for noise cancellation. Interestingly, the training accuracy of mass increased from 90.81% to the range of 93.13% - 96.44% after applying to pre-process and fine-tuning a pertained Alex-net; and the training accuracy of MCs increased from 87.23% to a range of 90.4% - 95.84%.

7.4 Discussion

The experimental results validate the significant importance of using pre-processing before training CNNs. All the selected pre-processing methods yielded better classification performance compared to not using any pre-processing. We also employed other techniques like fine-tuning, data Aug., BN, and dropout in our work to reduce overfitting. To the best of our knowledge, our study [5] is the first to demonstrate the effect of different pre-processing image filters on the accuracy of CNNs for classification of mammography images. In our future work, we will apply pre-processing filters to the training MG data-set to enhance the accuracy of the used CNN model.

Chapter 8

Convolutional Neural Network (Vanilla U-Net) for Automated Mass Segmentation in Mammography

8.1 Abstract

Automatic segmentation and localization of lesions in mammogram (MG) images are challenging even with employing advanced methods such as deep learning (DL) methods. We developed a new model (Vanilla U-Net) based on the architecture of the semantic segmentation U-Net model to precisely segment mass lesions in MG images. The proposed end-to-end Vanilla U-Net extracts contextual information by combining low-level and high-level features. We trained the proposed Vanilla model using huge publicly available databases, (CBIS-DDSM, BCDR-01, and INbreast), and a private database from the University of Connecticut Health Center (UCHC) called UCHCDM.

We compared the performance of the proposed Vanilla model with those of the state-of-the-art DL models including fully convolutional network (FCN), SegNet, Dilated-Net, and Faster R-CNN models and the conventional region growing (RG) model. The proposed Vanilla U-Net model outperforms the Faster R-CNN method significantly in terms of the runtime and the Intersection over Union metric (IOU). Training with digitized film-based and fully digitized MG images, the

proposed Vanilla U-Net model achieves a mean test accuracy of 99.7%. The proposed Vanilla model achieves a mean Dice coefficient index (DI) of 0.952 and a mean IOU of 0.934 that show how close the output segments are to the corresponding lesions in the ground truth maps. Data augmentation has been very effective in our experiments resulting in an increase in the mean DI and the mean IOU from 0.922 to 0.952 and 0.918 to 0.934, respectively.

The proposed Vanilla U-Net based model can be used for a precise segmentation of masses in MG images. This is because the segmentation process incorporates more multi-scale spatial context, and captures more local and global context to predict a precise pixel-wise segmentation map of an input full MG image. These detected maps can help radiologists in differentiating benign and malignant lesions depend on the lesion shapes. We show that using transfer learning, introducing augmentation, and modifying the architecture of the original model results in better performance in terms of the mean pixel accuracy, the mean DI, and the mean IOU in detecting mass lesions compared to the other DL and the conventional models.

8.2 Background

Breast cancer is the second most common cause of cancer death among women in the United States [199]. According to the American cancer society, the female breast cancer death rate declined by 38% from its maximum in 1989 to 2014 (avoiding about 300,000 deaths) [199]. In 2012, the estimated number of deaths among females in the USA is 43,909 out of 293,353 of all cancer deaths. Moreover, in 2017, it is estimated that there will be 40,610 breast cancer deaths in the USA [67,199]. This decline in mortality is partially due to the advances in mammography screening and conventional computer-aided diagnosis models (CAD) [205]. In the last few years, deep learning

(DL) models and, in particular, convolutional neural networks (CNNs) have achieved state-of-the-art performance for image classification, lesion detection for mammography [4, 127], and for medical applications in general [139]. Various approaches have been proposed to further improve the accuracy of deep CNNs [4].

In a recent survey [156] on conventional CAD models and DL classification models for mammograms (MGs) images, it has been shown that conventional models have limitations in classifying MG images. Most of the conventional models depend on a prerequisite set of local hand-crafted features that cannot be generalized to work on a new data-set. Conventional CAD models consider limited feature types (e.g. texture features, shape features, and grey level intensity features), which require expert knowledge for selecting them [156]. Poor feature extraction and selection cause challenge to build a successful classifier [4, 156] (see Fig. 8.1). However, the state-of-the-art CNNs, extract global features from MG images [4, 98]. In CNNs, the first layers of the network capture basic coarse features such as oriented edges, corners, textures, and lines while subsequent layers construct complex structures or global features [127] (see Fig. 8.2).

Despite the initial success of DL methods for the segmentation of lesions in medical images as general, the segmentation of lesions in mammography using DL methods has not been studied thoroughly. A few studies have used a CNN-based model for lesion segmentation [6, 219] in MGs and more research need to be done in this topic [4, 139]. Few studies have employed CNN-based models for lesion detection and localization [6, 17–19, 21, 22, 49, 56, 57, 124, 185, 212, 240, 257]. These detectors provide bounding boxes (BBs) indicating regions of interests (ROIs), not real lesion segments. The region-based CNN (R-CNN) models [81] and its faster variants, Fast R-CNN [80], and Faster R-CNN [183] have recently become more popular for localization tasks

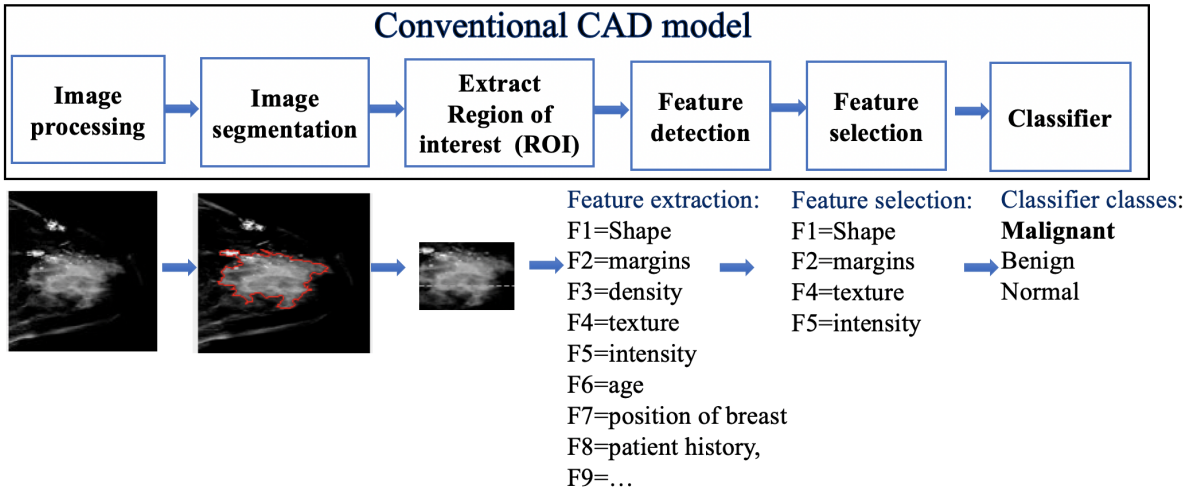


Fig. 8.1: Conventional CAD models consider limited feature types (e.g. texture features, shape features, and grey level intensity features),which require expert knowledge for selecting them.

in mammography [17, 18, 56, 57, 185]. Although these detectors offer compelling advantages, training R-CNN is time-consuming and memory expensive. In R-CNN [81], the whole process involves training three independent models separately without much-shared computation: 1- the CNN for feature extraction, 2- the top SVM classifier for identifying ROIs' and 3- the regression model for tightening region BBs. The R-CNN [81] uses the Selective Search method [221] to first generate initial sub-segmentations and generate candidate regions, then it uses the greedy algorithm to recursively combine similar regions into larger ones, and lastly uses the generated regions to produce the final candidate region proposals. These region proposals lower down the number of the potential BBs [17, 185].

Instead of extracting CNN feature vectors independently for each region proposal, the Fast R-CNN [80] aggregates them into one CNN forward pass over the entire image and the region proposals share this feature matrix. Then the same feature matrix is used for learning the object classifier and the BB regressor. In R-CNN and Fast R-CNN, the region proposals are created using

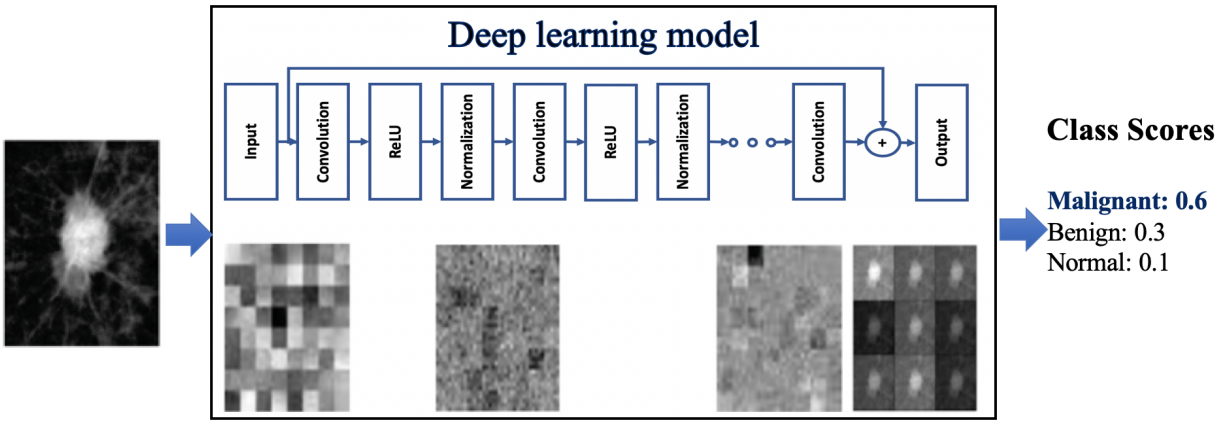


Fig. 8.2: The state-of-the-art CNNs extract global features from MG images, the first layers of the network capture basic coarse features such as oriented edges, corners, textures, and lines while subsequent layers construct complex structures or global features.

the Selective Search method, which is a slow process found to be the bottleneck of the overall detection and the localization process. The Faster R-CNN [183] is a better approach that constructs a single unified model composed of region proposal network (RPN) and Fast R-CNN with shared convolutional feature layers. The RPN is a fully convolutional CNN that is trained to generate region proposals, which are then used by the Fast R-CNN for detection. The time cost of generating region proposals is much smaller in the case of RPN than Selective Search, as RPN shares the most computation with the object detection network using the shared convolution layers [80, 183].

The mask R-CNN for simultaneously detecting and segmenting object instances in an image is proposed in [91]. This method extends the Faster R-CNN method by adding a branch which is a fully convolutional network (FCN) for predicting an object mask in parallel with the existing branch for BB recognition. A mass detector has been refined using a cascade of R-CNN and RF classifiers and adding an additional stage to eliminate false positives [56].

Patch-based CNNs [49, 55, 212, 240] were also proposed to detect masses. In [212], every breast

image is divided into patches, and each patch is tested with the CNN method individually. The final detection of lesions in each case is based on the overall scores of all the patches. In [19, 21, 22] the famous YOLO CNN (You Only Look Once) [181] is used for breast mass classification and localization. YOLO [181] is a single end-to-end CNN that predicts BBs and class probabilities directly from full images in one evaluation.

Recently, the FCN and its variant improved models as U-Net [186], SegNet [27], Dilated-Net [246], have yielded outstanding results for semantic segmentation of bio-medical images and natural images [4, 6, 98]. These semantic segmentation networks are based on encoding (convolutional) and decoding (de-convolutional) layers. These approaches avoid using the fully connected layers (FCLs) of CNNs to convert the image classification networks into image semantic segmentation networks.

In this study, we developed a new model based on the architecture of the semantic segmentation U-Net model [186] to precisely segment mass lesions in MG images. In the proposed Vanilla U-Net architecture, we used a pre-trained encoder layers and we added batch normalization layers (BN) [105], and dropout layers [208]. U-Net [186] is a complete end-to-end model that takes an image, find automated features in each layer, detects, and segments breast lesion using a single model and a unified training process. We trained the proposed Vanilla U-Net model using large public datasets (CBIS-DDSM [52], BCDR-01 [142], and INbreast [155]). We applied data augmentation (aug) to the training images to present the lesions in many different sizes, positions, angles. To enhance the contrast of the MGs, we applied image pre-processing before training the proposed Vanilla U-Net model. We compared the performance of the proposed Vanilla U-Net segmentation model in detecting lesions with those of the state-of-the-art Faster R-CNN [185], the conventional

region growing (RG) [148], FCN [141], Dilated-Net [246], and SegNet [27] models.

8.3 Material and Methods

8.3.1 Data-sets

We conducted our experiments on four Data-sets, CBIS-DDSM [52], INbreast [155], UCHC [254], and the BCDR-01 [142]. CBIS-DDSM [52] is a digitized screen-film mammography (SFM) database that is a subset of the digitized DDSM database [93] with updated lesion segmentation and BBs, and verified pathology. We used 1,696 images from the CBIS-DDSM database that have mass lesions. BCDR-D01 is an SFM repository with 64 patients and 246 MGs [142]. In total, we used 136 mass segmentation from this database to conduct our experiments. The INbreast is another public database for MGs which comprises fully field digital mammography (FFDM) images [155]. It has a total of 410 images, and we used 116 MGs that are annotated for masses. A database of FFDM images was collected from the University of Connecticut health center (UCHCDM) [254]. In total, the UCHC database consists of 173 patients with 1,340 FFDM images. We selected 59 cases out of the 173 that have mass lesions, with a total of 118 MGs with mass annotations.

We combined these databases and generated a new data-set containing MGs with different resolutions. This new data-set provides mass lesions of different sizes, shapes, and margins. All images containing suspicious areas have associated pixel-level ground truth maps (GTMs) indicating the true locations of suspicious regions. The total number of images used in this combined data-set is 2,066 and each image has its corresponding GTM. We divided the images into a training data-set of 1714 images, validation data-set of 204 images, and test data-set of 148 images. Images reserved for testing were not used in the training and the validation data-set. Images that come from the

same patient were not split across the training and test data-sets.

8.3.2 Data Pre-Processing

Pre-processing of MGs is an essential step before applying DL methods. Its main goal is to enhance the characteristics of MGs by applying a set of filters to improve the performance of the down-stream analysis. First, we detect the breast boundary for removing a big portion of the black background [78,178]. After that, we apply the adaptive median filter (AMF) [83] to remove any existing noises. Then, we employ the contrast limited adaptive histogram equalization (CLAHE) [260] to enhance the contrast of the MGs [4, 78, 178]. The superior performance of the CLAHE filter compared to other filters are shown in [4, 5]. The AMF [260] is a nonlinear filter that removes impulse noise while preserving edges and corners to improve the image quality. The CLAHE filter increases the contrast between the masses and their surrounding tissues [19, 41, 55, 57]. The CLAHE [260] filter operates on small regions in the image, called tiles, rather than the entire image. It calculates the contrast for each tile individually producing local histograms. Each tile's contrast is enhanced and the neighboring tiles are then combined using bilinear interpolation to eliminate artificially induced boundaries. The contrast in the homogeneous regions can be limited using a clipLimit factor to avoid amplifying any noise that might be present in the image. We used tiles of [8 8] and a clipLimit factor of 0.005 with the CLAHE technique. Figure 8.3 shows a sample of the combined data-set we used in our experiments. Figure 8.4 shows images containing suspicious areas and their associated pixel-level GTMs. All full MGs and GTMs are converted into png format and re-sized to 512×512.

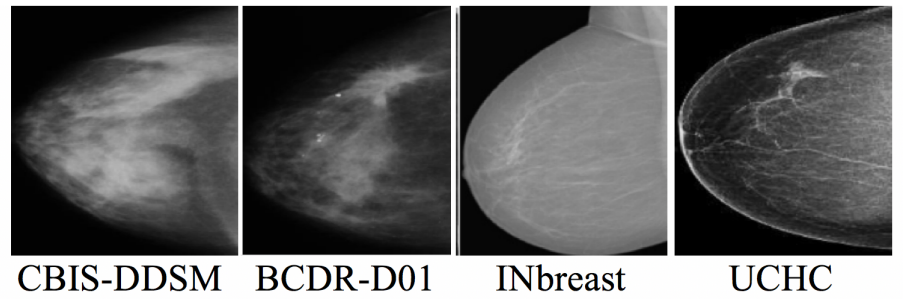


Fig. 8.3: The databases used in our experiments.

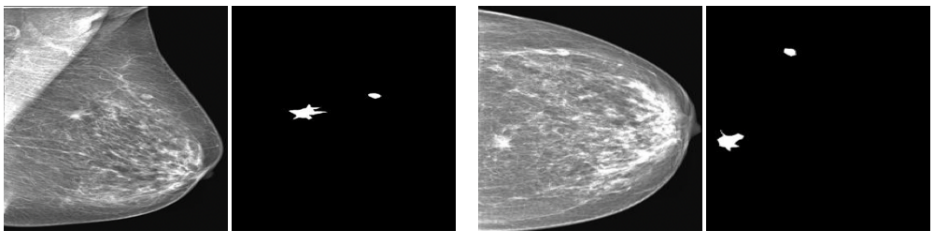


Fig. 8.4: MG images and their corresponding GTMs.

8.3.3 Data Augmentation

In this study, we adopt augmentation techniques to increase the size of our training data-set to avoid overfitting the model. We generated augmented images by image rotation in a range of ± 10 degrees, left-right flips, translate image left and right by 10%, translate images up and down by 10%, and zoom in and out by 20%. The mass segmentation maps are represented by binary images that are cropped, re-sized and augmented in the same way as their corresponding MGs. All pixels in the GTMs are labeled as belonging to background or breast lesion classes. The size of the generated augmented data-set is ten times larger than the size of the original data-set.

8.3.4 Semantic Segmentation using U-Net

The U-Net is a popular end-to-end encoder-decoder network for semantic segmentation that is originally invented for bio-medical image segmentation tasks [186]. U-Net [186] extends the FCN [141] with a U-shape architecture, which allows features from shallower layers to combine with those from deeper layers. U-Net consists of a contracting path to capture features and an asymmetric expanding path that enables precise localization and segmentation of pixels. This architecture has a U shaped skipping connections that connect the high-resolution features from the contracting path to the up-sampled outputs of expanding path. After collecting the required features in the encoding path, the decoding path performs nonlinear up-sampling of the feature maps before merging with the skip connections from the encoding path followed by two 3×3 convolutions, each followed by an element-wise rectified linear unit (ReLU). The skip concatenation allows the decoder at each stage to learn back relevant features that are lost when pooled in the encoder. The final output is obtained by passing the result through a pixel-wise Softmax classifier after the last convolution layer, which independently assigns a probability to each pixel.

8.3.5 Proposed Architectural Modifications to U-Net

We have modified the original U-Net model [186] to improve its performance for the task of segmenting lesions. We added BN layers [105], dropout layers [208], and increased the number of convolution layers. We also trained the model with augmented data-set and used a pre-initialized encoder layer with weights from a pre-trained VGG16 model [201]. BN [105] prevents internal covariate shifts as data are filtered through the network, and it reduces the training time, prevents data overfitting, helps stack more layers, and generally increases the performance of deep CNNs. We

added drop-out layers of 0.5 after each convolutional layer to help regularize the networks [208]. Figure 8.5 show our modified model. The encoding path consists of five convolutional layers which perform convolution with a filter bank to produce a set of feature maps. A BN layer is added between the convolution layer and the ReLU layer. Following that, max-pooling with a 2×2 window and stride 2 is performed and the resulting output is sub-sampled by a factor of 2. The max-pooling layer reduces the dimensionality of the resulting output, enabling the further collection of features. Also, to keep the size of the output map the same as the size of the original input MGs, a padded convolution is applied to keep the dimensions consistent across concatenation levels. In this work, we pre-initialized the encoder layers of the proposed Vanilla U-Net model with weights from a pre-trained VGG16 model.

Our data-set has imbalanced data representation. In an imbalanced representation, classes are represented by significantly different numbers of pixels, which makes the learning algorithm biased towards the dominating class (i.e. breast tissues and/or background). We address this problem by introducing class weights into the Dice loss function [86]. The class weight is the ratio of the median of class frequencies computed on the entire training set divided by the class frequency [86]. This implies that the breast tissues and background class in the training set have weights smaller than the weights of the lesion class. Moreover, we applied the augmentation techniques explained in the previous sub-section, instead of applying elastic deformations as done in the original U-Net method [186].

For training, the Dice loss function was minimized using Adam optimizer [118] with a decreasing learning rate (LR) initialized to $1e^{-}2$ and a momentum of 0.9. We trained the networks for 100 and 150 epochs. Before each epoch, the training set is shuffled and each 4 mini-batch images

are then picked thus ensuring that each image is used only once in an epoch. We used input MGs re-sized to 512×512. We developed, trained, and tested the DL models using MATLAB version 2017b. Training and testing the models were done on a Tesla K40m Nvidia graphics processing unit.

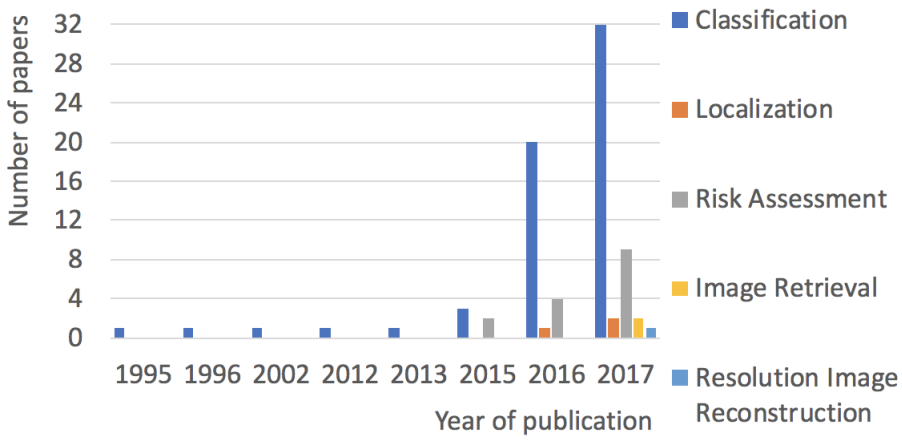


Fig. 8.5: The U-Net architecture consists of convolutional encoding and decoding units that take an image as input and produce the segmentation feature maps with respective pixel classes. The yellow arrows show the skip connections between the layers.

8.3.6 Semantic Segmentation using FCN

Semantic segmentation is an active research area for medical images where deep CNNs are used to classify each pixel in the image individually. Semantic segmentation results in a map image that is segmented by classes. The FCN [141] is an encoder-decoder network. The encoder path uses a pre-trained VGG16 model [201] and transfer their learned representations by fine-tuning to the segmentation task. The decoder path uses up-sampling operations, and replace the final FCL with an N×1×1 convolution layer, which output probabilities for N classes. A skip architecture is proposed by [141] where the weights of shallow, fine layer features are combined with deep, coarse

layer features to produce accurate and detailed segmentations, as intensive up-sampling can lead to coarse segmentation maps. There are 3 versions of FCN (FCN-32s, FCN-16s, FCN-8s) based on VGG16 network [141]. In this research, we adapt the FCN-8s VGG16 based network [141] to our segmentation task. FCN-8s up-samples the final feature map by a factor of 8 after fusing feature maps from the third and fourth max-pooling layers.

8.3.7 Semantic Segmentation using SegNet

The SegNet architecture [27] adopts the VGG16 network [201] along with an encoder-decoder framework wherein it drops the FCLs of the network. SegNet shares a similar architecture to the encoder-decoder U-Net described in the previous subsection. However, in SegNet, the indices at each max-pooling layer in the encoder contracting path at each level are stored and later used to up-sample the corresponding feature map in the decoder by unpooling it using those stored indices (Fig. 8.6). Storing the indices from the contraction path helps keep the high-frequency information intact, however, it also misses neighboring information when unpooling from low-resolution feature maps. Finally, a Softmax classifier is used to produce the final segmentation maps with the same resolution of the original MG image. In this work, we used a SegNet that is pre-initialized with layers and weights from a pre-trained VGG16 model with an encoder D of 5.

8.3.8 Semantic Segmentation using Dilated-Net

Recently, Dilated-Net [246], also known as atrous convolutions, have been used in different image segmentation tasks [47,158,226,237,247]. Dilated convolutions [246] allow us to explicitly control the resolution at which feature responses are computed and incorporate a larger context without increasing the number of parameters or the amount of computation. We adopt the dilated CNN

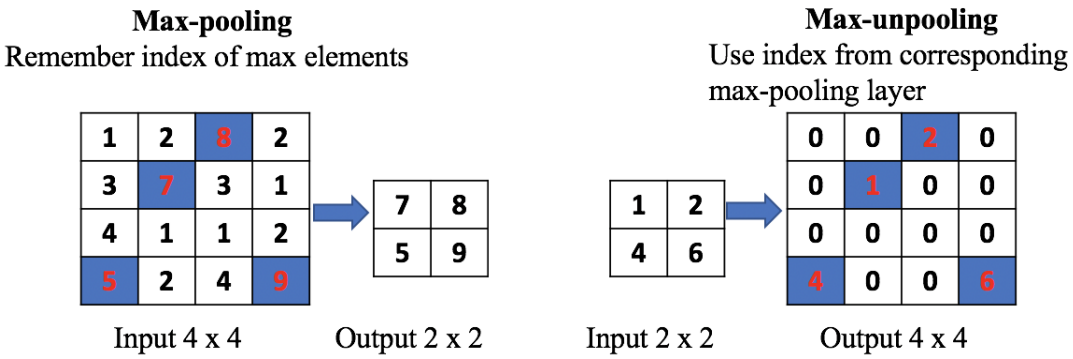


Fig. 8.6: In SegNet, the indices at each max-pooling layer in the encoder contracting path at each level are stored and later used to up-sample the corresponding feature map in the decoder by unpooling it using those stored indices.

in [246] with some modifications to the network. The implemented dilated CNN architecture consists of ten cascaded 3×3 convolutional layers with dilation factors 1, 1, 2, 4, 8, 16, 32, 1, 1 and 1 (Fig. 8.7). Figure 8.7 illustrates a 3×3 convolution kernels with different dilation factor as 1,2, and 3. The last three layers are FCLs of 1×1 convolutions followed by dropout of 0.5 [208]. The first nine convolutional layers are followed by BN layer [105] and a ReLU activation function [157]. To classify the pixels, the last convolutional layer has two 1×1 convolutions, followed by a Softmax classifier.

8.3.9 Localization using Faster R-CNN

We adapt the Faster R-CNN method proposed in [185] to compare its performance in terms of accuracy of detection and inference time with that of the proposed Vanilla U-Net model. Faster R-CNN is based on a VGG16 model [201] with additional components for detecting, localizing and classifying lesions in MG image. Faster R-CNN outputs a BB for each detected lesion, and a score, which reflects the confidence in the class of the lesion. The Faster R-CNN method in [185]

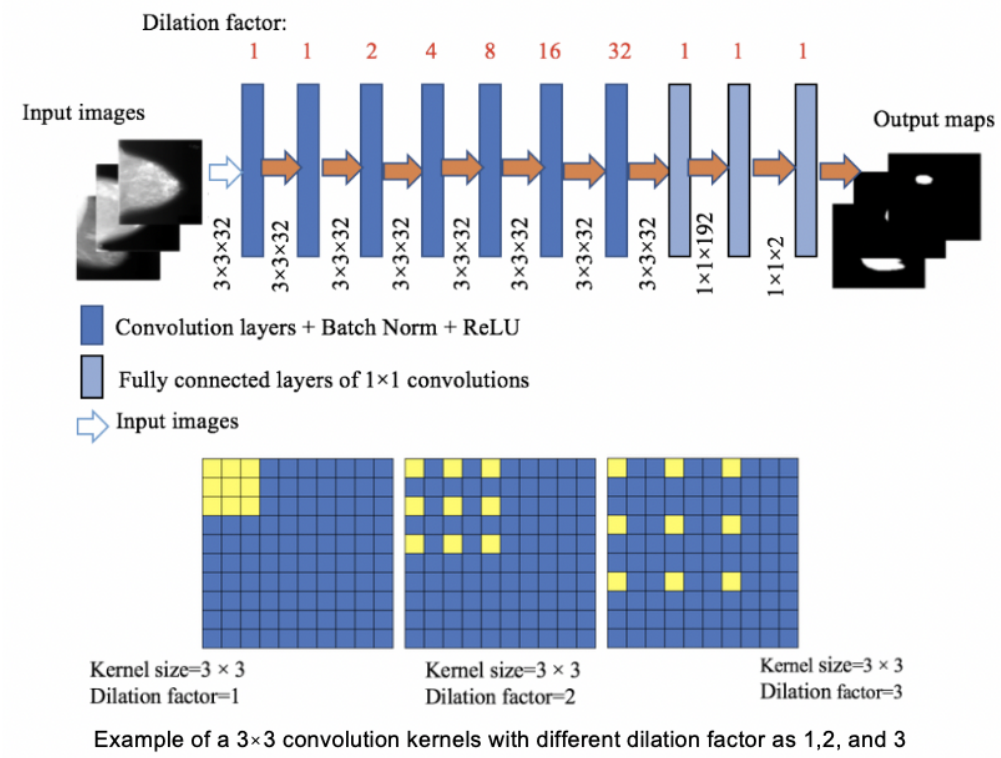


Fig. 8.7: Architecture of the dilated-Net, containing ten convolutional layers with dilation factors, indicated in red, increasing from 1 in the first layer to 32 in the seventh layer. The last 1×1 convolutional layer is followed by a Softmax classifier.

is trained with our pre-processed and augmented data-set. Further details about the implemented Faster R-CNN method can be found in the original article in [185]. One limitation stated in the study of [185] is that the training data-set comes from small-sized publicly available pixel-level annotated data-set. However, in our study, we are using our combined large-sized data-set to re-produce their work.

8.3.10 Semantic Segmentation using Region Growing (RG)

We also implemented the region growing (RG) model proposed in [148] and apply it to our MG images. RG is a traditional image segmentation CAD model that starts with selecting an initial seed

point and then groups pixels or sub-regions into larger regions according to a similarity criterion. As RG results are sensitive to the initial seeds, the automated accurate seed selection is very critical for image segmentation. Further details about the implemented RG method can be found in the original article [148].

8.3.11 Comparison between State-of-the-art DL Methods in the Literature

Table 8.1 lists the information about the architecture, databases, the number of images, the evaluation methods (i.e. Accuracy (ACC.), area under the curve (AUC), Dice index (DI)), TPR@FPR, the testing time per image as provided in the literature.

Table 8.1: Shows a comparison between the proposed segmentation method and the current state-of-the-art DL methods for segmentation or localization of lesions in MG images.

Paper	Approach	Tested on	#Images	Size	TL	Aug.	AUC	ACC.	DI	TPR@FPR	Inference time/image
Proposed Vanilla U-Net	U-Net	INbreast, DDSM, UCHC	17,140	512×512	Yes	Yes	-	99.7	95.23	-	0.12s
Rilbi et al. [185]	Faster R-CNN	INbreast	3,927	1400×1700	Yes	Yes	95.0	N/A	N/A	0.90@0.3	N/A
Choukroun et al. [49]	Batch-based CNN	INbreast	410	224×224	Yes	Yes	72.2	N/A	N/A	N/A	N/A
	INbreast	2,500	224×224	Yes	Yes	83.1	N/A	N/A	N/A	N/A	
Akselrod et al. [17]	Faster R-CNN	In-house	850	800×800	Yes	Yes	72.0	77.0	N/A	0.93@0.56	0.2s
Akselrod et al. [18]	Faster R-CNN	INbreast	100	224×224	Yes	Yes	N/A	N/A	N/A	0.93@0.56	5s
		INbreast	310	224×224	Yes	Yes	N/A	N/A	N/A	0.8@1.5	5s
Dhungel et al. [56]	R-CNN+ Random forest (RF)	INbreast	410	40×40	Yes	Yes	N/A	95.0	N/A	0.91@1.3	41s
Dhungel et al. [57]	R-CNN+ RF	INbreast	410	40×40	Yes	Yes	91.0	95.0	85	0.9@1.0	41s
Plataniaet al. [169]	YOLO	DDSM	10,480	128×128	Yes	Yes	N/A	90.0	N/A	N/A	1.25s
Teuwen et al. [219]	Faster R-CNN	In-house	23,405	227×227	Yes	Yes	N/A	N/A	N/A	0.87@0.56	N/A
	Mask R-CNN	In-house	23,405	227×227	Yes	Yes	N/A	N/A	N/A	0.97@3.5	N/A
Al-masni et al. [21]	YOLO	DDSM	600	448×448	No	No	87.74	96.33	N/A	0.96@2.6	N/A
Al-masni et al. [22]	YOLO	DDSM	600	448×448	No	Yes	96.45	99.70	N/A	N/A	3s
Sun et al. [212]	Batch-based CNN	In-house	840	52×52	No	Yes	69.82	N/A	N/A	N/A	N/A
Reiazi et al. [182]	Faster R-CNN	MIAS	120	75×75	No	No	N/A	N/A	N/A	N/A	N/A

8.4 Evaluation Metrics

To evaluate the performance of the DL networks, the Dice index coefficient (DI), also known as the F1 score, and the Intersection over Union (IOU), also known as the Jaccard index, metrics are used to compare the automated predicted maps with the GTMs [53, 72]. We mapped the class probabilities from the Softmax output to discrete class labels and used it to calculate the commonly used DI and IOU metrics, Eqs. (8.1) and (8.2) respectively.

$$DI = \frac{2 \times TP}{2 \times TP + FP + FN}, \tag{8.1}$$

where TP is the number of true positive pixels, FP is the number of false positives and FN is the number of false negatives.

$$IOU = \frac{TP}{TP + FP + FN}, \tag{8.2}$$

For each class, IOU is the ratio of correctly classified pixels to the total number of ground truth and predicted pixels in that class. The average IOU of each class is weighted by the number of pixels in that class. We use this metric as the images have dis-proportionally sized classes, to reduce the impact of errors in the small classes on the aggregate quality score.

As mentioned in the Background section, most of the lesion detection models provide BBs for an indication of a region with an abnormality. To compare the performance of the proposed Vanilla U-Net model with detection methods providing BBs such as the Faster R-CNN, a BB is generated around every detected lesion. The BBs are generated based on a minimum and maximum points of x and y coordinates, which indicate the locations of masses. We calculated the accuracy of localization by considering the detected segment and BB as TP if the center of the segment or the BB overlaps with the ground truth by more than 50%. For each class, the pixel accuracy metric is the ratio of correctly classified pixels to the total number of pixels in that class, according to the GTMs, (Eq. 8.3). Mean pixel accuracy is the average accuracy of all classes in all images. We also calculated the global accuracy which is the ratio of correctly classified pixels, regardless of class, to the total number of pixels.

$$Pixel\ accuracy = \frac{TP}{TP + FN}. \tag{8.3}$$

We also calculated the Boundary F1 contour matching score (BF-score) for each image, which indicates how well the predicted boundary of each class aligns with the true boundary. For each class, mean BF-score shows the average BF-score of all classes in all images. Values near 1 means

perfect boundary.

8.5 Segmentation Results

8.5.1 Comparison with State-of-the-art Methods

We adopted the Faster R-CNN model [185], VGG16-based FCN-8s model [141], VGG16-based SegNet model [27], Dilated-Net [246] model, and the conventional RG CAD model [148] to apply to MGs for comparing their performances with that of our model in terms of mean pixel accuracy, mean DI, mean IOU, mean BF-score, and the inference time in seconds per image. We trained the Faster R-CNN detector proposed in [185] to detect breast cancer lesions on MGs using our augmented data-set. We also implemented the RG model proposed in [148] and apply it to our MG images.

The test data-set consists of SFM and FFDM MG images. Figures 8.8 (b) and 8.9 (b), show the original FFDM MG images from the INbreast database and the SFM MG images from the DDSM database, respectively. Where the red BBs show the ground truth given by radiologists. In Figs 8.8 and 8.9, (c) shows the GTMs, (d) shows the output of the FCN method trained for 150 epochs on the augmented data-set, (e) shows the output of the Dilated-Net method trained for 150 epochs on the augmented data-set, (f) shows the output of the SegNet method trained for 150 epochs on the augmented data-set, (g) shows the output of the proposed Vanilla U-Net method trained for 150 epochs without augmentation, (h) shows the output of the proposed Vanilla U-Net method trained for 150 epochs on the augmented data-set, (i) shows the output of the RG method, (j) shows the output of the Faster R-CNN method, and finally (k) shows the green BBs surrounding the detected

tumors using the proposed Vanilla U-Net method trained for 150 epochs on the augmented data-set. The calculated IOU for each detection is shown under each image. Tables 8.2 and 8.3, show the evaluation metrics of all the networks included in this study.

Table 8.2: The performance of the FCN, Dilated-Net, SegNet, RG, and the proposed Vanilla U-Net model.

Model	**Mean pixel acc**	**Mean IOU**	**Mean BFscore**	**Mean DI**	**Mean time (sec)**
-Proposed Vanilla U-Net, Aug., 150 epochs.	0.997	0.934	0.966	0.952	0.116
-Proposed Vanilla U-Net, Aug., 100 epochs.	0.994	0.918	0.965	0.922	0.096
-Proposed Vanilla U-Net, no Aug., 150 epochs.	0.988	0.918	0.955	0.922	0.102
-Proposed Vanilla U-Net, no Aug., 100 epochs.	0.986	0.917	0.947	0.848	0.101
-Original U-Net, no Aug., 150 epochs.	0.954	0.842	0.899	0.818	0.118
-SegNet (Aug., 150 epochs)	0.988	0.914	0.961	0.925	0.097
-SegNet, no Aug., 150 epochs.	0.978	0.908	0.960	0.921	0.093
-Dilated-Net, Aug., 150 epochs.	0.957	0.903	0.945	0.840	0.093
-Dilated-Net, no Aug., 150 epochs.	0.948	0.892	0.938	0.837	0.093
-FCN, Aug., 150 epochs.	0.957	0.912	0.938	0.847	0.090
-FCN, no Aug., 150 epochs.	0.947	0.906	0.904	0.841	0.091
-RG	0.853	0.755	0.701	0.789	0.3

Table 8.3: The performance of the FCN, Dilated-Net, SegNet, Faster R-CNN, RG, and the Proposed Vanilla U-Net model.

Model	Mean pixel acc.	Mean IOU	Mean inference time (sec) /image
-Proposed Vanilla U-Net, Aug., 150 epochs.	0.925	0.902	0.116
-Original U-Net, no Aug., 150 epochs.	0.853	0.812	0.118
-Faster R-CNN, Aug., 150 epochs.	0.900	0.855	0.453
-SegNet, Aug., 150 epochs.	0.872	0.823	0.097
-Dilated-Net, Aug., 150 epochs	0.883	0.842	0.093
-FCN, Aug., 150 epochs	0.876	0.872	0.090
-RG	0.811	0.788	0.320

Table 8.4: The performance of the Proposed Vanilla U-Net method.

Model	Training		Validation		Test
	Dice	Loss	Dice	Loss	Dice
-Proposed Vanilla U-Net, with Aug.	0.994	0.014	0.972	0.021	0.952
-Proposed Vanilla U-Net, no Aug.	0.924	0.062	0.863	0.146	0.848

In Table 8.2, the performance of the models is shown for the detected segments in comparison

with the GTMs. In Table 8.3, the performance of the models under study is shown for comparison between the tight detected BBs and the ground truth BBs. The mean DI and the mean IOU of the proposed Vanilla U-Net are 0.952 and 0.934, respectively, which are higher compared to other methods (Table 8.2). The BF-score of the proposed Vanilla U-Net method is 0.966 which exceeds the other segmentation models. The architecture of the SegNet method is much closer to that of the U-Net method compared to the other segmentation methods. However, the boundary of the detected regions of SegNet model is not aligned with the true boundary (Figs. 8.8 (f) and 8.9 (f)). The SegNet model has a BF-score of 0.961. The SegNet method performs better when detecting lesions in FFDM MGs compared to SFM MGs. In contrast, the proposed Vanilla U-Net model performs very well for both kinds of images (Figs. 8.8 (h) and 8.9 (h)). The proposed Vanilla U-Net shows better performance compared to SegNet. U-Net transfers the entire feature maps to the corresponding decoders and concatenates them to the up-sampled decoder feature maps, which gives precise segmentation. SegNet has much fewer trainable parameters compared to the U-Net model since the decoder layers use max-pooling indices from corresponding encoder layers to perform sparse upsampling. This reduces the inference time at the decoder expanding path since the generated encoder feature maps are not involved in the upsampling. Thus, the SegNet method reveals a trade-off between the memory verses mean pixel accuracy involved in achieving good segmentation performance (Tables 8.2 and 8.3). The mean DI and the IOU of the best trained SegNet model are 0.925 and 0.914, respectively, compared to 0.952 and 0.934 of the U-Net model.

The trained Dilated-Net has a mean DI of 0.840, a mean IOU of 0.903, respectively. Moreover, its BF-score is 0.841 that is lower than that of the proposed Vanilla U-Net model and the SegNet model BF-score. Also, the performance of the Dilated-Net model is worse in the case of SFM

images (Fig. 8.8 (e)). Even-though some images in Figs. 8.8 (e) and 8.9 (e) show slightly better IOU than that of SegNet, the performance of the model on all the test data-set is lower than that of the SegNet model. In contrast to U-Net and SegNet, down-sampling layers are not required in the Dilated-Net to obtain large receptive fields and hence, high-resolution maps can be directly predicted by the model. Down-sampling layers are widely used for maintaining invariance and controlling overfitting of the model, however it reduces the spatial resolution. To retrieve the lost spatial information, the Up-sampling layers in U-Net, and SegNet are used, but with additional memory and time constraints.

We also adapted the FCN-8s VGG16 based network [141] to compare its performance with that of the proposed Vanilla U-Net model. FCN-8s up-samples the final feature map by a factor of 8 after fusing feature maps from the third and fourth max-pooling layers, thus having better segmentation than its variants. The FCN in our study has a mean DI of 0.847 and a mean IOU of 0.912, respectively. Moreover, the BF-score of the best trained FCN is 0.938 which is lower than that of the proposed Vanilla U-Net by 0.0282. The mean DI scores of the Dilated-Net and the FCN model are close for some of the images, however, the FCN give the lowest scores among all the segmentation DL methods.

As we mentioned before, we generated BBs surrounding detected segments to compare the performance of the proposed Vanilla method with that the BB-based methods such as Faster R-CNN. The proposed Vanilla U-Net model shows better performance in detecting true segments compared to the Faster R-CNN model as shown in Fig. 8.8 (j: k) and Fig. 8.9 (j: k). In Fig. 8.8, the Faster R-CNN model introduces FPs in the SFM images as in rows (1j), (4j: 5j), and (7j: 9j). In Fig. 8.9 (j), the Faster R-CNN method introduces some FPs as in row (8j) and missed some lesions

as in row (5j), as an example. The proposed Vanilla U-Net method shows better performance with both FFDM and SFM images than the Faster R-CNN method. To have a better understanding of the performance of the proposed Vanilla U-Net and other methods, we included the IOU under every image. The IOU of the proposed Vanilla method exceeds the IOU of the Faster R-CNN by 0.047 (Table 8.3). We considered the detected BBs as TP if the center of the detected BB overlaps with the ground truth BB with greater than 50%.

The accurate automated seed selection process is very important for lesion segmentation. As RG segmentation's results are sensitive to the initial seed pixels, the final segmentation results would be incorrect if the seeds are not properly selected by the automated process. The RG method works better when it is used with patches of images that contain the ROI because the initial seed pixels are close to the center of the ROI. Figures 8.8 (i) and 8.9 (i) show the detection using the RG method. Figures 8.8 and 8.9, show that the DL methods outperform the CAD models in terms of IOU in the segmentation of tumors in whole images. The mean DI and mean IOU of the RG method are 0.789 and 0.755, respectively.

8.5.2 Effect of Augmentation and Increasing the Epochs

We investigated the effect of augmentation in the performance of the proposed Vanilla U-Net method. The augmented training data-set results in 17,140 images. Figures 8.8 (g: h) and 8.9 (g: h), illustrate the effect of augmentation on the proposed Vanilla U-Net method. For example, the values of the DI of the augmented model, a shown in (h), are higher than the ones of the trained model without augmentation, as shown in (g). Table 8.4 shows the improvement in terms of DI for both training and validation data-sets when using augmented data-set compared to when using the

original one. The DI improves from 0.924 (training), and 0.863 (validation) to 0.994 (training), and 0.972 (validation). The augmented data also affect the localization precision significantly (Table 8.2). The BF-score improves from 0.947 to 0.966 in the case of the proposed augmented U-Net model.

Figure. 8.10 shows that the histogram of the mean of IOU value for the test images increases using the proposed Vanilla U-Net method after data augmentation. The mean of IOUs of the proposed Vanilla U-Net improves from 0.917 to 0.934 when training with the augmented data-set (Fig. 8.10 and Table 8.2). In general, the performance of the DL techniques improves as the size of the training data-set increases [4]. Figures 8.8 (g:h) and 8.9 (g:h), show that the IOU per image increases when the proposed Vanilla model is trained with the augmented mixed data-set. The FP pixels decreases in the case of augmented model, as shown rows (1,6,8,9) in Fig. 8.8 (g:h). We also trained the same network without data augmentation for 100 and 150 epochs to examine the effect of increasing the number of epochs on the segmentation of MGs using the proposed Vanilla U-Net. Tables 8.2 and 8.3, show that the DI for the test MGs increases with the increase of the number of training epochs.

8.5.3 Effect of Image Size and Data-set Size

One of the factors that make a localization model or a semantic segmentation model superior to other models, is its ability to help the radiologists to detect small lesions that can be missed with the naked eye. A recent study in [79] on MGs shows that the resolution of the training images affects the performance of the CNN model. Recent studies, as shown in Table 1 in our work in [8], use MGs of small sizes as 40×40 and 227×227. The standard image sizes of 224×224, and 227

$\times$ 227 are used excessively for training CNNs to detect objects in natural images [72]. However, the requirement to find small mass lesions in aggressively down-sampled high-resolution images is unlikely to be successful for MGs [4, 79]. In our initial work in [6], we trained the proposed Vanilla model with images of size 256×256 and found that the proposed Vanilla method failed to find small lesions in images of high density. As a result, we changed our training strategy to MGs of size 512×512 instead of 256×256. Figure 8.11 shows some FFDM test images that have small lesions that are detected with DI greater than 50%. In the future, we will conduct our experiments on high-resolution images to get a competitive performance to recent state-of-the-art models.

8.5.4 Improvements of the proposed Vanilla Model over the Original U-Net Model

The proposed Vanilla model yields an improvement of 16.32% in the DI and 10.89% in the IOU, respectively, relative to that of the original U-Net model (Table 8.2). The original U-Net model is trained from scratch. However, the proposed Vanilla model is pre-trained using natural images, then fine-tuned with MG images. Moreover, increasing the data-set size by using the proposed Vanilla augmentation technique improves the segmentation's quality (BF-score yields an increase of 7.46% relative to that of the original U-Net model). The original U-Net did not use the BN technique. BN helps the proposed Vanilla model avoiding vanishing gradient problem, stacking more layers, accelerating training, and using less number of epochs. In the proposed Vanilla model, we went deeper into the number of layers from four to five convolution layers. By increasing the number of convolution layers, the segmentation process incorporates a more multi-scale spatial context and captures a more local and global context.

8.5.5 Timing Performance

To assess the runtime performance of these methods, we measured the mean inference time per image taken by each method to detect lesions in the test data-set (Table 8.3). The proposed Vanilla U-Net method is faster by 0.3376 sec than the Faster R-CNN method [185]. The inference time of the SegNet, Dilated-Net, and FCN is less than the proposed Vanilla U-Net by a fraction of second. Even though the inference time of the RG method is of about 0.3 sec, it introduces a lot of FPs when tested on whole images as shown in Figs. 8.8 (i) and 8.9 (i), and the statistics of Tables 8.2 and 8.3. The proposed Vanilla U-Net method is faster than the Faster R-CNN proposed in [18] and [17], the R-CNN proposed in [56] and [57], and the Yolo method proposed in [22], while proving a high DI. We have to emphasize that for radiologists, the accuracy of the proposed CAD or DL model in detecting lesions is the most important feature in the mammography analysis, and the inference time is secondary. An inference time of a fraction of second or even several seconds is not as important as the accuracy of the given model.

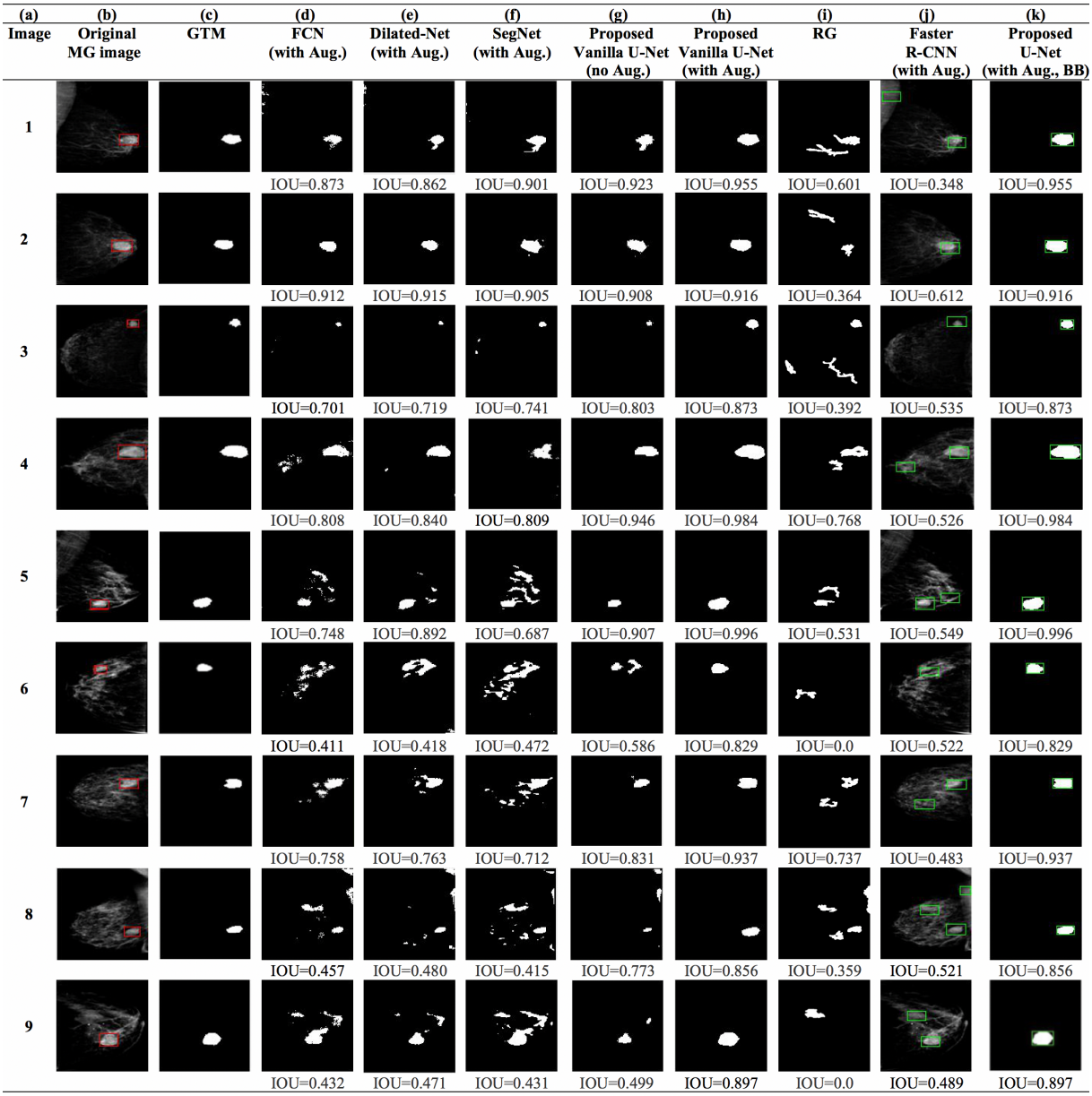


Fig. 8.8: (a) Image index, (b) the SFM MG images from the DDSM database (the red rectangles show the location or the BBs of the ground truth lesions), (c) the GTMs given by radiologist, (d) the FCN method, (e) the Dilated-Net method, (f) the SegNet method, (g) the proposed Vanilla U-Net method without augmentation, (h) the proposed Vanilla U-Net method trained with the augmented data-set, (i) output of the RG method, (j) the Faster R-CNN method, and finally (k) shows the green BBs surrounding the detected tumors using the proposed Vanilla U-Net method trained with the augmented data-set.

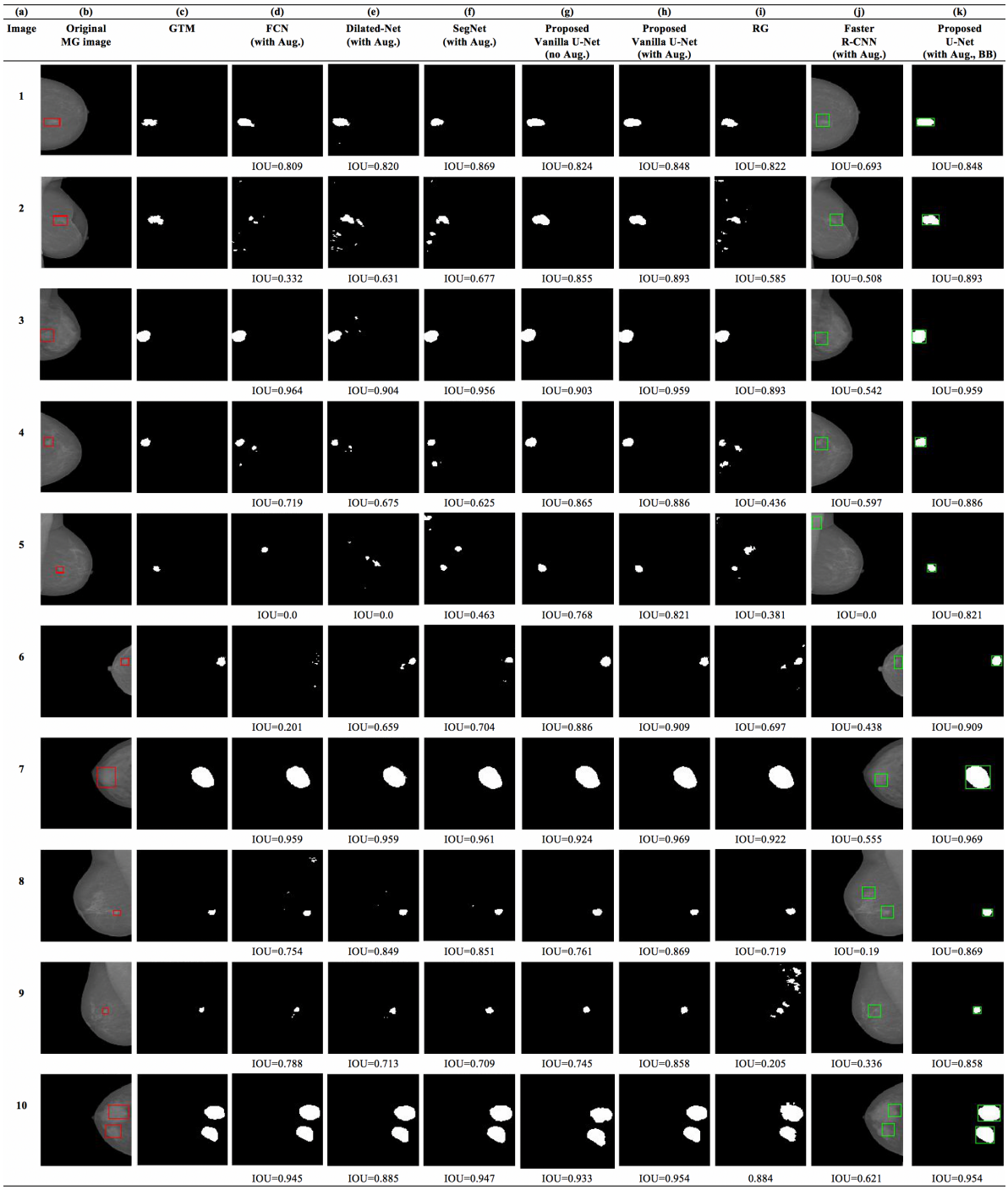


Fig. 8.9: Detection of lesions using different models. (b) shows the FFDM MG images from the INbreast database.

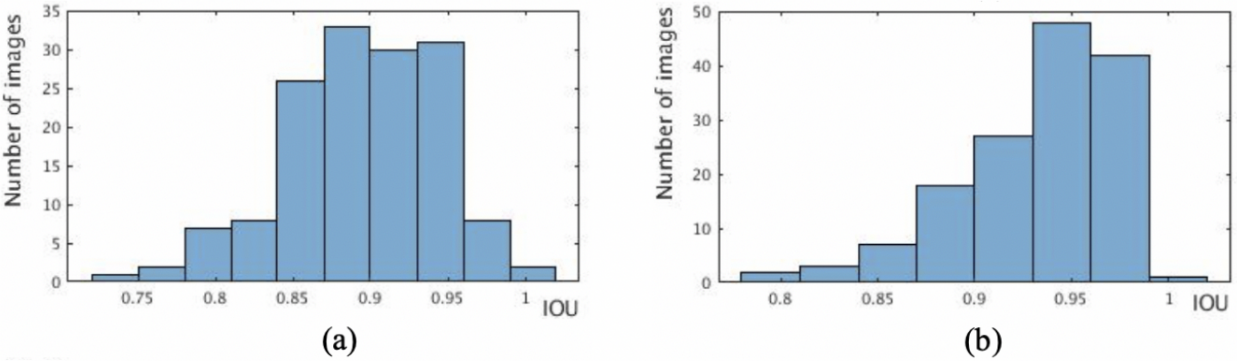


Fig. 8.10: Histogram of the mean of IOU value for the test images using the proposed Vanilla U-Net method before Aug. (a) and after Aug. (b).

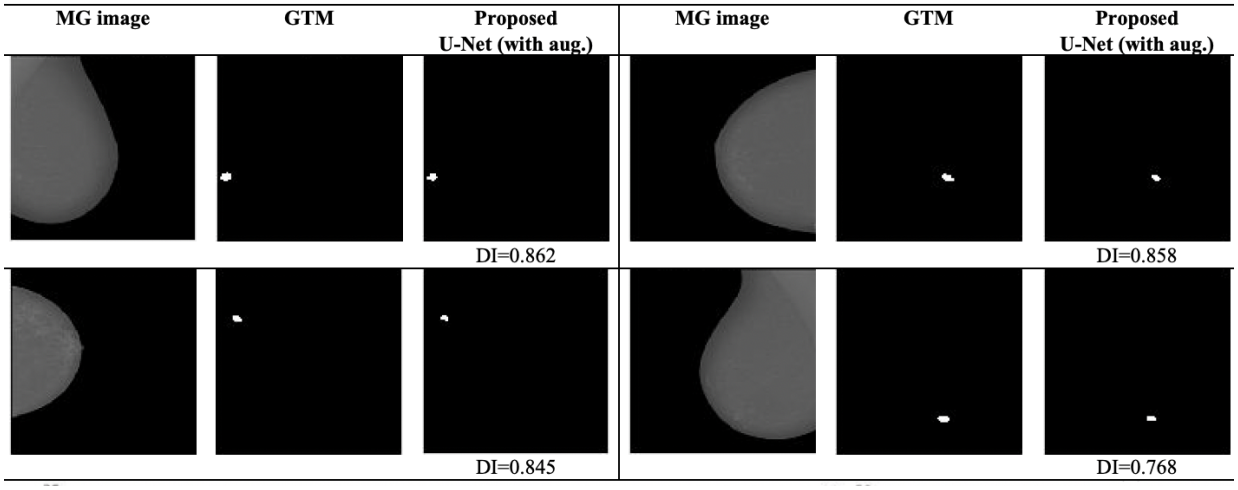


Fig. 8.11: Test cases with small lesions.

8.6 Discussion

We tested our proposed Vanilla method on SFM and FFDM data-sets for the semantic segmentation of mass lesions in MGs. For our future work, we will consider training the U-Net model to detect both the MCs and the masses lesions. We will focus on reducing FP pixels by collecting more data-sets and use higher resolution MG images. Finally, we want to use the U-Net to distinguish between benign and malignant breast tumors in mammography images by studying the features of the tumors' segmented regions only.

8.7 Conclusions

We developed a new model based on the architecture of the semantic segmentation U-Net model to precisely segment mass lesions in MGs. The proposed end-to-end based model extracts low-level and high-level features from MGs. The proposed Vanilla U-Net efficiently predicts a pixel-wise segmentation map of an input full MG due to its modified architecture. We tested our proposed Vanilla model using film-based and fully digitized MGs. We compared the performance of our model with state-of-the-art deep learning models namely Faster R-CNN, SegNet, FCN, and Dilated-CNN. We also compared the performance of the proposed Vanilla model with the conventional region growing model. The proposed Vanilla U-Net method is superior to the segmentation methods under study. We generated BBs surrounding the segmented lesions using U-Net to compare them to the detected BBs using the Faster R-CNN method. The proposed Vanilla U-Net model gives mean IOU of 0.903 and mean pixel accuracy of 0.926 while the Faster R-CNN model gives IOU of 0.855 and mean pixel accuracy of 0.901, respectively. Similar to the Faster R-CNN method, the U-Net method is trained on the full MGs. However, the proposed Vanilla U-Net method is faster

and runs 0.337 sec less than the Faster R-CNN method. We show that the proposed Vanilla method show improvement in the DI and the IOU by 16.32% and 10.89%, respectively, relative to the original model. The proposed methods can be further trained to detect micro-calcification in the future. The presented work is a step towards a precise segmentation of mass lesions in mammography. As medical data-sets are increasing and becoming publicly available, future architectures may be trained end-to-end, removing the need for pre-training on non-medical data-sets.

Chapter 9

Residual Deep Learning System for Mass Segmentation and Classification in Mammography

9.1 Abstract

Automatic extraction of breast mass in MG images is a challenging task due to the varying sizes, shapes, and textures of masses. Moreover, the density of MGs makes mass detection very challenging since masses can be hidden in dense MGs. In this work, we propose a residual DL system for mass segmentation and classification in mammography. The overall proposed system consists of two cascaded parts: 1) a residual attention U-Net model (RAttU-Net) to precisely segment mass lesions in MG images, followed by 2) a ResNet classifier to classify the detected binary segmented lesions into benign or malignant. The proposed semantic-based CNN model, RAttU-Net, has the basic architecture of the U-Net model, which extracts contextual information combining low-level features with high-level ones. We have modified the U-Net structure by adding residual attention modules to preserve the spatial and context information, help the network have deeper architecture, and handles the gradient vanishing problem. We compared the performance of the proposed RAttU-Net model with those of state-of-the-art two semantic segmentation models, and two object detectors using public databases. We also examined the effect of breast density on the accuracy

of localizing and segmenting the breast masses. Our proposed model shows superior performance compared to the other DL methods in detecting and segmenting masses, especially for heterogeneously dense and dense MG images, in terms of IOU and the DI. Moreover, our results show that the cascaded ResNet model, trained using binary-scale images, classify the masses to cancer and benign with higher accuracy compared to the ResNet model that is trained on gray-scale images.

9.2 Background

Mammograms (MGs) have contributed significantly to the reduction of the breast cancer mortality rate through early detection of cancer. Recent advances in computational technologies and significant progress in deep learning (DL) [127,180] and image processing techniques [161] have opened up unprecedented opportunities to develop models for providing an objective view to radiologists with higher accuracy [4,9,259]. With advances in detection and localization methods in DL techniques for medical imaging [139], few studies have proposed DL models to localize mass lesions in MG images [6,19,49]. Studies in [4,9,49] show that convolution neural networks (CNNs) achieve higher detection accuracy in locating masses in MGs compared to traditional Computer-Aided Detection (CAD) models. Various approaches have been proposed to further improve the accuracy of deep CNNs in detecting and localizing breast abnormalities [4,9,259]. Further, techniques such as stochastic depth [100], batch normalization (BN) [105], transfer learning (TL) [164], data augmentation (Aug.) [127], and dropout [208] have been used in various researches for avoiding network overfitting and regularization purposes. Despite the recent advances in the structure of DL models, detection of masses in MG images has remained a challenging problem due to the following reasons: 1) existence of some masses in the pectoral muscle area, 2) hidden masses under the dense

breast tissues, and 3) varying sizes, shapes, and texture of masses [4, 9, 259].

In this study, we propose a residual DL system for mass segmentation and classification in mammography. The overall proposed system consists of two cascaded parts: 1) a residual attention U-Net model (RAttU-Net) to precisely segment mass lesions in MG images, followed by 2) a cascaded ResNet [92] classifier to classify the detected binary segmented lesions into benign or malignant. The proposed semantic-based CNN model, RAttU-Net, has the basic architecture of the U-Net model [186], which extracts contextual information combining low-level features with high-level ones. In our previous work, we have shown that the basic U-Net model can be used for more precise and efficient mass segmentation in MG images [6]. To further improve the performance of the basic U-Net model for mass segmentation, we have modified its structure by adding residual attention modules. These modules generate attention-aware features that change adaptively as the network goes deep in layers. The residual modules [92] resolve the problem of vanishing gradients using identity skip-connections thus facilitating the training of the proposed model. The proposed RAttU-Net uses long and short skip connections to produce precise and detailed segmentation maps. Besides adding the stacked residual attention modules, we used augmented data-set in the training process to improve the accuracy of the RAttU-Net model.

To evaluate the performance of the proposed model, we compared the performance of the proposed RAttU-Net model with those of the basic U-Net model [186] and the vanilla U-Net model [6]. The performance of the models is evaluated in terms of dice index coefficient (DI), intersection over union (IOU) and inference test time. We also constructed bounding boxes (BBs) surrounding the segmented lesions to compare the output BBs produced by the proposed RAttU-Net with the BBs produced by the state-of-the-art Faster R-CNN and Yolo object detectors pro-

posed in [21,185]. We used publicly available data-sets, CBIS-DDSM [93], and BCDR-D01 [142] to train the proposed RAttU-Net model and the models we used for performance comparison proposed in [6,21,185,186], and we tested the models using the INbreast data-set [38]. These data-sets include mass lesions of different sizes, shapes, and margins. We also examined the effect of breast density on the accuracy of localizing and segmenting the breast masses. We tested the performance of our proposed RAttU-Net and the models in [6, 21, 185, 186] separately on each breast density category based on the BI-RADS code: fatty, scattered, heterogeneously dense and dense breasts, in terms of the Dice index coefficient (DI) and the intersection over-index (IOU). RAttU-Net is implemented in Matlab R2018b and is available at https://github.com/NabaviLab/RAttU-Net.

9.3 Related Work

Despite the huge success of DL methods in classifying MG images into normal, benign and malignant [4,9,259], the use of CNNs for segmentation and localization of lesions in MG images are not thoroughly investigated. The region-based CNN (R-CNN) models and its faster variants, Fast R-CNN, Faster R-CNN, and Mask R-CNN [80,91,183] for object detection have recently been used for mass localization tasks in mammography [182,185,219]. Yolo (You Only Look Once) that is an effective and efficient object detection DL model has been used for mass localization [19,21,22].

Anchor boxes are widely adopted in state-of-the-art DL object detection models (e.g. Faster R-CNN and Yolo V2). The main drawback of these models is that if anchor boxes are not chosen correctly, the model will struggle in detecting small or irregular objects. In the Yolo model, the anchor shapes are obtained by k-means clustering on the sizes of the ground truth BBs. In Faster R-CNN, the anchor shapes are of 3 scales and of 3 aspect ratios, yielding 9 different anchors at

each output sliding window position. The aspect ratios used in the case of detecting general objects such as pedestrian, car, text is different from the aspect ratios used in detecting lesions in MGs, or medical images in general. In these models, the anchor shape has to be manually modified according to the ground-truth data-set to improve the detection accuracy and to detect small mass lesions. The Faster R-CNN [183] and Yolo [181] models need to pre-define the anchor box's shapes and to fix its size during training, which is sub-optimal since it ignores the augmented data distribution in training. Inappropriate anchor boxes could degrade the performance of the detector in terms of accuracy [250] and detecting small lesions.

Recently, the fully convolutional network (FCN) and its variant improved models such as U-Net [186] and SegNet [27], have yielded outstanding results for semantic segmentation of biomedical images and a promising results for segmenting lesions in MG images [6, 257]. In these studies [19, 22], to enhance the detection performance of the used Yolo model in terms of precision in detecting masses, the authors first, used a cascaded semantic CNN to segment the detected masses. Then, they used another cascaded CNN trained on gray-scale MG images to classify the segmented masses as either benign or malignant [19, 22]. These cascaded models increase the computation cost, however, they provide better classification results.

Several studies [49, 240] have proposed a patch-based CNN to detect lesions. The drawback of the patch-based approaches is that the input patches came from non-overlapping areas, which makes it difficult to precisely localize masses. Moreover, the size of the input patches is very small that produces difficulty in differentiating normal tissues from abnormal ones after detection.

In this study, we propose a residual DL system to segment and classify lesions in MG images without the need for any cascaded detectors. To date, only a few attempts based on DL have been

presented for semantic segmentation of mass lesions in mammography [6, 257].

9.4 Material and Methods

9.4.1 Data-sets

We trained our proposed model on two data-sets, DDSM [93] and BCDR-01 [142], and tested it on the INbreast data-set [38]. All images containing masses have associated pixel-level boundary of the mass lesions. To have ground truth for evaluating object detection methods, we generated ground truth BBs for the masses based on minimum and maximum points values of x and y coordinates of the mass' contours, which indicate the locations of masses. Each MG image has been annotated based on their density derived from the American College of Radiology's (ACR) Breast Imaging Reporting and Data System (BI-RADS) [29]. For each MG image, its density in ACR standard scale is given as one of these categories: class A: fatty, class B: scattered, class C: heterogeneous dense, and class D: dense [4, 29]. We grouped the BI-RADS multi-class assessment into benign and malignant classes. In this study, we categorized 1133 MG images with BI-RADS $\in$ {2, 3} as benign, and 1843 MG images with BI-RADS $\in$ {4, 5, 6} as malignant. The distribution of density for each BI-RADS class is presented in Table 10.2. In total, we used 2,976 MG images to conduct our experiments.

9.4.2 Data Pre-Processing

We first detect the breast boundary for removing a big portion of the black background [78] from the training images. After that, we employ the CLAHE to enhance the contrast of the MG images.

Table 9.1: Distribution of breast density in each BI-RADS class in the publicly available data-sets used in our study.

Purpose	Data-set	BI-RADS	Class				Total
			A	B	C	D	All
Training	**DDSM**	Benign	54	215	228	224	1,018
Training	**BCDR-D01**	Benign	34	16	23	7	80
Test	**INbreast**	Benign	12	4	13	6	35
Training	**DDSM**	Malignant	219	693	513	291	1,716
Training	**BCDR-D01**	Malignant	16	15	22	2	55
Test	**INbreast**	Malignant	30	32	8	2	72

We generated GTMs for the masses using the associated pixel-level boundary of the mass lesions given by the data-sets. All pixels in the GTM are labeled as belonging to the background (0) or breast lesion (255) classes. All full MGs images and it's corresponding GTMs are re-sized to 640 $\times$ 640. To deal with the small training data-set and avoiding overfitting our model, we applied data augmentation to the training MG images and it's corresponding GTMs by image rotation by (-45, 45) degrees, translation up and down by (-10%, 10%), scaling in and out by 0.2, and left-right flips.

9.4.3 Proposed RAttU-Net

U-Net is a popular E2E encoder-decoder network for semantic segmentation that is originally invented for bio-medical image segmentation [186]. U-Net consists of a contracting path to capture features and an asymmetric expanding path that enables precise localization and segmentation

of pixels. This architecture has a U-shaped skipping structure that connects the high-resolution features from the contracting path to the up-sampled outputs of expanding path. Inspired by the residual attention mechanism proposed in [229], we built the proposed RAttU-Net model by stacking residual attention modules to the basic U-Net architecture. We use the residual blocks with the identity connections instead of the regular convolution layers in the U-Net architecture to preserve the spatial and context information, help the network have deeper architecture, and handles the gradient vanishing problem. The residual blocks directly propagate features from its early convolution to its late convolution and improve the performance of the proposed model consequently. To address the problem of detecting small lesions, the proposed RAttU-Net model uses residual attention blocks to increase the resolution for better pixel-level prediction (Table 9.2).

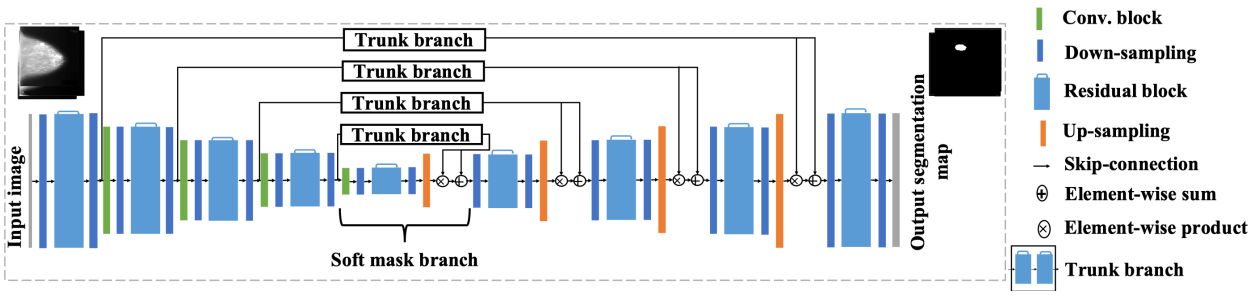


Fig. 9.1: The architecture of the proposed RAttU-Net. The nested long residual skip connections connect the encoder and decoder paths at the same level, while each intermediate residual block contains a short residual skip connection within the same path to increase the depth of the proposed RAttU-Net model.

The residual attention module consists of a soft mask branch and trunk branch [229] (Fig. 9.1). The attention residual mechanism can keep the flow of original feature information through the trunk branch using the identity mapping and construct attention to mass lesions features using the soft mask branch. Each trunk branch is connected to its soft mask branch (Fig. 9.1).

(a) Proposed RU-Net. (b) Vanilla U-Net.

Fig. 9.2: Difference between the proposed RAttU-Net (a) and the vanilla U-Net (b). The proposed RAttU-Net model is a fully residual model that has long and short skip connections.

(a) A convolution block. (b) A residual block.

Fig. 9.3: A convolution block consists of a BN layer, ReLU layer, and convolutional layer (a) and a residual block consists of the element-wise sum of the output of three cascaded convolutional blocks with the identity map to produce the final feature output (b).

The proposed RAttU-Net network consists of multiple levels, and in each level, the network capture features with different resolutions. As shown in Fig. 9.2, the encoder in the soft mask branch at each level consists of a cascade of a down-sampling layer, a convolution (Conv.) block, a residual block, another convolution block, and a skip connection of residual blocks that is connected to the corresponding level in the decoder path. The attention module keeps useful information by applying element-wise product between feature coming from the truck branch and the output of the soft mask branch. An element-wise sum is then performed between the output of the element-wise product and output from the residual blocks in the trunk branch. This element-wise summation relieves the feature attenuation that happened during the element-wise product process

Table 9.2: Architecture of the proposed RA U-Net. The symbol ↓ means that this level in the encoder path consists of a convolution block, a residual block, a convolution block, and a down-sampling layer. The symbol ↑ means that this level in the decoder path consists of a convolution block, a residual block, a convolution block, and a down-sampling layer. The symbol □ means that this level consists of a bridge of convolution block, a residual block, and another convolution block.

Layer name	**Path**	**Layers inside**	**Output resolution**	**Output width**
Input	Encoder	↓	640×640	1
Level 1	Encoder	↓	640×640	64
Level 2	Encoder	↓	320×320	128
Level 3	Encoder	↓	160×160	256
Level 4	Encoder	↓	80×80	512
Bridge		□	40×40	1024
Level 4	Decoder	↑	80×80	512
Level 3	Decoder	↑	160×160	256
Level 2	Decoder	↑	320×320	128
Level 1	Decoder	↑	640×640	64
Classifier	Decoder	↑	640×640	1

by using long connections, which enhances the feature contrast. The output from the element-wise summation is then forwarded to the decoder path, which at each level consists of a cascade of a

convolution block, a residual block, another convolution layer, and finally an up-sampling deconvolution layer. The final feature output of every residual block is the element-wise summations of the output of three cascaded convolutional blocks with the short identity map (Fig. 9.3). Each convolutional block consists of a BN layer and an activation ReLU layer and a regular convolution layer (Fig. 9.3). The down-sampling (max-pooling) layers exist between the levels in the encoding path to perform down-sampling in the feature maps. The deconvolution layers exist between levels in the decoding path to up-sample the input feature maps from the decoder level and then concatenate them using a pixel-wise addition with the feature maps coming from the encoding path by the long skip connections.

Besides the long skip connections used between each level in the encoder-decoder path, short connections are used in the residual blocks for a direct connection between layers in the same levels. Using short and long connections help the flow of information within and across levels in the RAttU-Net architecture to generate richer information hierarchy (Fig. 9.1). The trunk branch in each level uses its long skip connection as input to a cascade of two residual blocks (Fig. 9.1). The output of the truck branch is then element-wise summed with the up-sampling feature maps from the corresponding level in the decoder path. The final segmented binary map is obtained by passing the result through a pixel-wise Sofmax classifier after the last convolution layer.

The original U-Net model uses concatenation [186] of feature maps between the encoder and decoder path. In this work, the concatenation is replaced by element-wise summation (Fig. 9.2). Element-wise summation directly adds the local details of the feature maps from the encoder to the global details of the feature maps from the decoder at a certain stage. Thus, the residual attention modules generate attention-aware features that change adaptively as the network goes deep in

layers.

The novelty of this work is that the residual attention modules are added to the vanilla U-Net model [6], as in Fig. 9.3, to capture multi-scale information and integrate low-level features with high-level features for precise semantic segmentation of the input MG images. By using the residual attention modules, our RAttU-Net can improve the performance significantly (Fig. 9.1). We investigated employing a different number of residual attention blocks, and a different number of layers. We observed that adding more than two residual blocks and four layers do not significantly improve the model's performance, but it significantly increases the training time. Therefore we considered the architecture shown in Table 9.2 for the proposed RAttU-Net model.

9.4.4 Cascaded Classifier

The classification of benign and malignant mass lesions is one of the most challenging and also the most significant processes in examining MG images as it helps to reduce false positives (FPs) and classify the lesions at their early stage. Almost all the DL methods proposed for classifying MG images use the gray-scale images [9]. We used images of 224×224 pixels to train our DL model for the classification of lesions in MG images into benign or malignant. The size 224×224 is used excessively for training DL CNNs [4, 9]. Classifying lesions into benign or malignant using gray-scale images is challenging because, in some MG images, gray levels of the masses are mixed with surrounding tissues resulting in unclear lesion boundary. The augmented training binary-scale GTMs have clear mass boundaries and margins. Moreover, heterogeneously dense and dense breast tissues hide the mass lesions. These challenges decrease the accuracy of DL classifiers. In this work, to address these challenges, we propose to use a cascaded ResNet CNN [92]

trained on black and white images to classify the segmented binary maps (output of the semantic segmentation) into benign or malignant images without the need of the traditional hand-crafted features (Fig. 9.4).

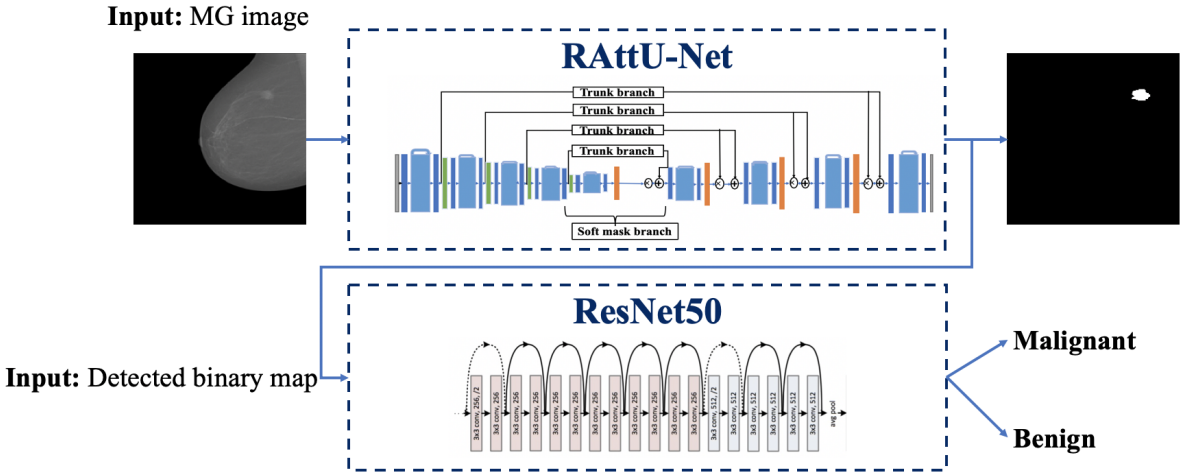


Fig. 9.4: Block diagram of the proposed cascaded modules for semantic segmentation and classification of mass lesions.

Classification of Binary-scale Image

In this work, to address the above challenges, we propose to use a cascaded ResNet CNN [92] trained on black and white images to classify the segmented binary maps (output of the semantic segmentation) into benign or malignant images without the need of the traditional hand-crafted features.

We used a pre-trained Res-Net model trained on natural images and then, fine-tune it with binary-scale GTMs by modifying the last fully connected layers to fit our task of binary classification. We used GTMs of the DDSM and BCDR-01 data-sets to train the model into benign and malignant images. We trained the models with 1098 benign MG images, and 1771 malignant MG

images, respectively (Table 10.2).

Classification of Gray-scale Mammogram

To perform classification on large MG images, the common approach used in the literature is to first develop a classifier to recognize smaller patches, and then use the classifier to scan the whole image using a sliding window. This is followed by another process to summarize the patch classifier's outputs to give the final classification result. However, in this research, we want to utilize a patch-based classifier to initialize the weights of a whole image classification network.

We developed an E2E training algorithm for whole-mammogram images to classify it to benign or malignant. The model requires lesion annotations only at the first stage of training. After that, a whole image classifier can be trained using only image-level labels. We trained a patch-based ResNet CNN model to classify ROIs into normal, benign and malignant patches. Each ResNet block has a shortcut that is made between the two ends of each unit so that the features are directly carried over and therefore each unit can focus on learning the "residual" information. BN is used in every convolutional layer in the ResNet, which is known to speed up convergence and also has a regularization effect. We added on top of the pre-trained patch ResNet new residual blocks to turn the patch network itself into a network that can detect whole MG images. The top residual blocks are followed by an average pooling layer and then the image's classification output. To make this new network detect and classify abnormalities on the whole image, we just need to modify the input of the new network from patches to whole images. Once the patch classifier is converted into a whole image classifier, it can be fine-tuned on other databases using only image-level labels. This way, instead of generating one prediction for a single patch, the network will generate predictions for all overlapping patches in the MG image, creating a likelihood probability of the MG being

benign or malignant.

9.4.5 Network Configurations

For training the segmentation models, we adopted the Dice coefficient (DC) loss [149] as the objective function to train the model. The DC loss function is minimized using Adam optimizer with a decreasing learning rate (LR) initialized to 10^{-2} and a momentum of 0.9. The LR is reduced every 25 epochs by a factor of 0.1. We trained the models for 150 epochs. We trained the models using mini-batches of size 4. To manage imbalance data, we introduced class weights into the DC loss function.

We utilized the ResNet CNN model [92] in this study for the classification of masses into benign or malignant. In our experiment, the ResNet model [92] is trained using a Stochastic Gradient Descent with a gamma of 0.1, and a weight-decay of 10^{-5}. We trained the models using mini-batches of size 16. We used initial LR of 10^{-3}, which is reduced every 25 epochs by a factor of 0.1. A dropout of 0.5 is used to accelerate the training process and prevent overfitting. For training the classification models, we used GTMs of the DDSM and BCDR-01 data-sets to train the model into benign or malignant images. We trained the models with 1,098 benign MG images, and 1,771 malignant MG images, respectively, as shown in Table 10.2. We applied the same augmentation technique discussed in the Material and Method Section to augment the binary GTMs images. We performed data augmentation to alleviate the relatively small amount of training data-set. To evaluate the performance of the ResNet model for classification, we carried out 5-fold cross-validation tests on the INbreast data-set (Table 10.2). The detected segmented binary images that have IOU $\geq$ 0.4 with the GTMs are used to test the cascaded classifier. The detected segmented binary images are resized to 224×224 pixels. We used a pre-trained Res-Net model trained on natural images

and then, fine-tune it with binary-scale GTMs by modifying the last fully connected layers to fit our task of binary classification. We also fine-tuned the pre-trained ResNet model using gray-scale MG images to classify MG images to benign or malignant. Both ResNet models were trained under the same settings and the same augmentation technique. We developed and trained the DL algorithms using MATLAB version 2018b. Training and testing the models were done on a Tesla K40m Nvidia graphics processing unit.

9.4.6 Evaluation Metrics

To evaluate the performance of the DL models, the DI, also known as the F1 score, and the IOU, also known as the Jaccard index, metrics are used to compare the automated predicted maps with the GTMs [73]. We mapped the class probabilities from the Softmax output to discrete class labels and used them to compute the DI and IOU metrics. As mentioned in the Related Work Section, most of the lesion detection models provide BBs for an indication of a region with an abnormality. To compare the performance of the proposed RAttU-Net model with object detection models that provide BBs, such as the Faster R-CNN and Yolo, we generated a BB around every detected lesion or segment. We used the minimum and the maximum points of x and y coordinates, which indicate the locations of masses to generate the BBs. We considered a detected segment (or BB) as true positive (TP) if it overlaps with the ground truth segment (or BB) by more than 40%. For each class, the pixel accuracy metric is the ratio of correctly classified pixels to the total number of pixels in that class, according to the GTMs. Mean pixel accuracy is the mean accuracy of all classes in all images. We also calculated the Boundary F1 contour matching score (BF-score) for each image, which indicates how well the predicted boundary of each class aligns with the true boundary. For each class, the mean BF-score shows the mean BF-score of all classes in all images, where values

near 1 show perfect boundary.

9.5 Segmentation Results

We compared the performance of the original U-Net model [186], Vanilla U-Net [6], Faster R-CNN model [185], and Yolo model [21] in detecting masses with that of our proposed model in terms of mean DI, mean IOU, and the inference time in seconds per image. Further details about the implementation of these models can be found in the original article in [6, 21, 185, 186]. We trained the above models using our augmented data-set and tested them using the INbreast data-set [38]. The outputs of the Faster R-CNN and the Yolo models are BBs per inference, as shown in Fig. 9.5 (h: i) in red. These models provide multiple BBs. Overlapping BBs are merged when the IOU between two boxes exceeds 0.5. The BB with the highest confidence score among the set of overlapping inference results is used as the representative BB with its confidence score. Figure 9.5a shows the original FFDM MG images from the INbreast data-set [38], where the green BBs show the location of lesions as given by radiologists, (b) shows the associated pixel-level GTMs of the mass lesions, (c) shows the prediction of the original U-Net model [186], (d) shows the prediction of the vanilla U-Net model [6], (e) shows the prediction of the proposed RAttU-Net model trained for 150 epochs without augmentation, (f) shows the prediction of the proposed RAttU-Net model trained for 150 epochs on the augmented data-set, (g) shows the output of the proposed best RAttU-Net model trained for 200 epochs on the augmented data-set, (h) shows the prediction of the Faster R-CNN model, and finally (i) shows the prediction of the Yolo model. In Fig. 9.5g, we constructed red BBs that surround the predicted masses using the proposed RAttU-Net method to compare them with the Faster R-CNN and Yolo detected BBs. The calculated DI and/or IOU for each

prediction is shown under each image in Fig. 9.5. The DI number shows the Dice similarity coefficient between the current predicted map and its corresponding GTM. DI takes a value in the range [0, 1], where 1 means that the segmentation in the two images is a perfect match. We also labeled each MG image with the corresponding breast density class (Fig. 9.5a). We added these labels to visually find the effect of breast density on the detection of masses in MGs across different models. Classes A, B, C, and D are corresponding to fatty, scattered, heterogeneously dense, and dense classes, respectively.

Table 9.3: The performance of the vanilla U-Net, Faster R-CNN, Yolo, and the proposed RAttU-Net model.

Model	**Class A**		**Class B**		**Class C**		**Class D**	
	Mean Acc.	**Mean IOU**	**Mean Acc.**	**Mean IOU**	**Mean Acc.**	**Mean IOU**	**Mean Acc.**	**Mean IOU**
-Proposed RAttU-Net	0.997	0.981	0.929	0.937	0.995	0.930	0.940	0.928
-Vanilla U-Net	0.921	0.873	0.931	0.912	0.912	0.911	0.939	0.912
-Faster R-CNN	0.763	0.717	0.799	0.732	0.733	0.689	0.881	0.797
-Yolo	0.886	0.814	0.878	0.791	0.911	0.778	0.910	0.828

The proposed RAttU-Net outperforms the segmentation results of the vanilla U-Net model [6] and the original U-Net model [186] in terms of DI (Fig. 9.5). Its better performance is due to the superiority of its network architecture. In the vanilla and original U-Net models, the aggregations from the encoder to decoder consist of simple and linear skip connections. As a result, the high-resolution features aggregated to the decoder are relatively shallow. Even-thought the vanilla

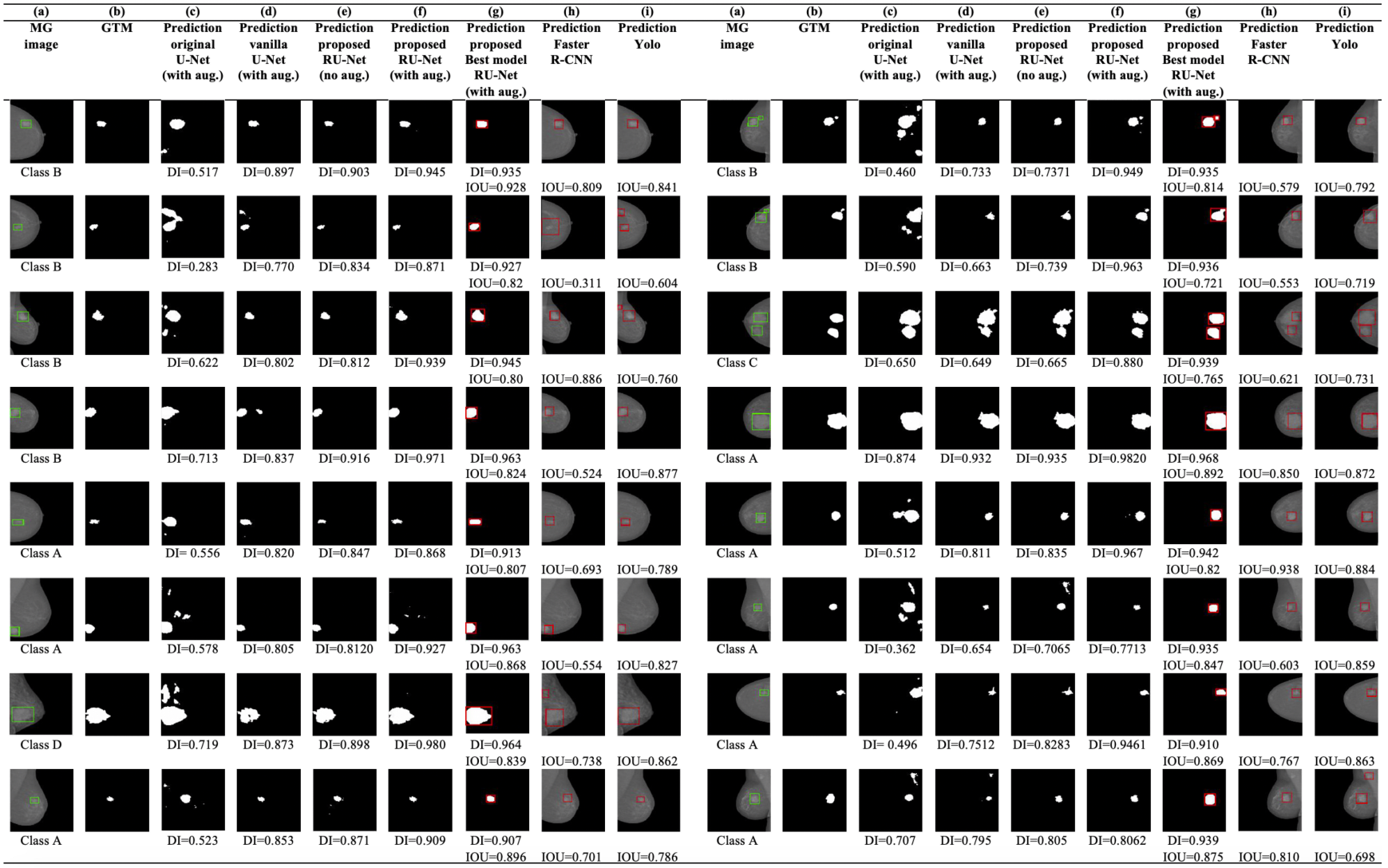


Fig. 9.5: A comparison between the original U-Net model, vanilla U-Net model, Faster R-CNN model, Yolo model, and the proposed RAttU-Net model for detection of mass lesions in the INbreast database.

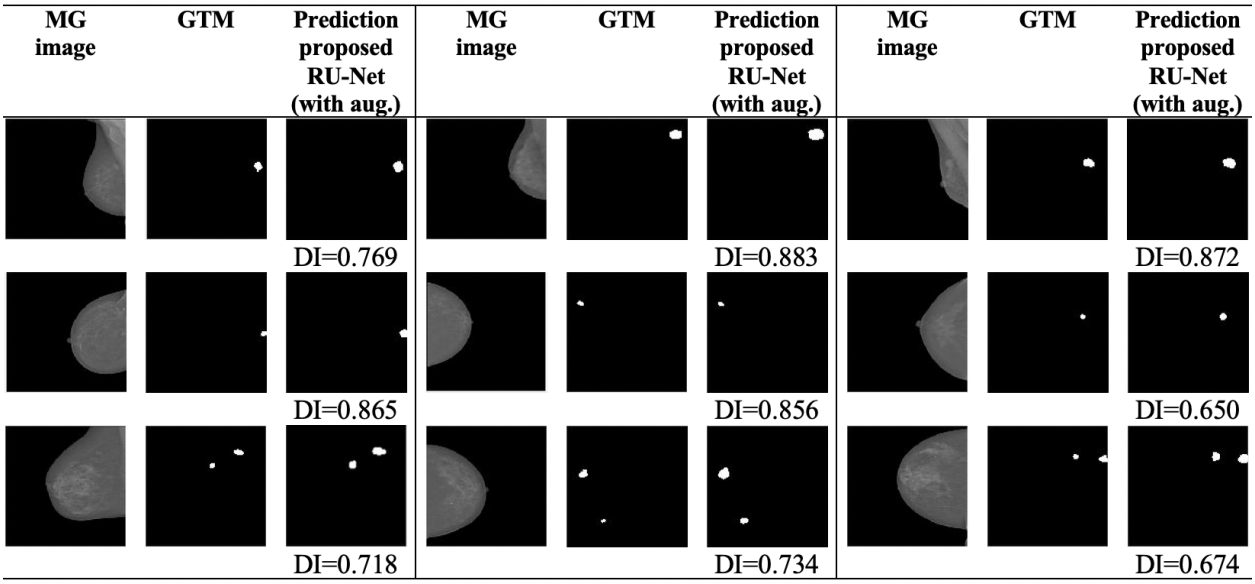


Fig. 9.6: Shows the detection of small lesions in the pectoral muscle, small lesions in heterogeneous dense MG images, and multiple small lesions.

Table 9.4: The performance of the proposed RAttU-Net without Aug., vanilla U-Net with Aug., original U-Net with Aug., and the proposed RAttU-Net model with Aug.

Model	Mean Acc.	Mean IOU	Mean BF-Score	Mean Dice
-Proposed RAttU-Net, with aug.	0.987	0.948	0.981	0.983
-Proposed RAttU-Net, without aug.	0.944	0.891	0.919	0.905
-Vanilla U-Net [6]	0.962	0.921	0.926	0.943
-Original U-Net [186]	0.843	0.836	0.789	0.756

U-Net model gives better results than the original U-Net in terms of DI, the proposed RAttU-Net gives more precise segmentation results. The FP segments in some of the scattered MG images in Fig. 9.5 (c: d) disappeared when using the proposed RAttU-Net. Figure 9.5g shows high IOU

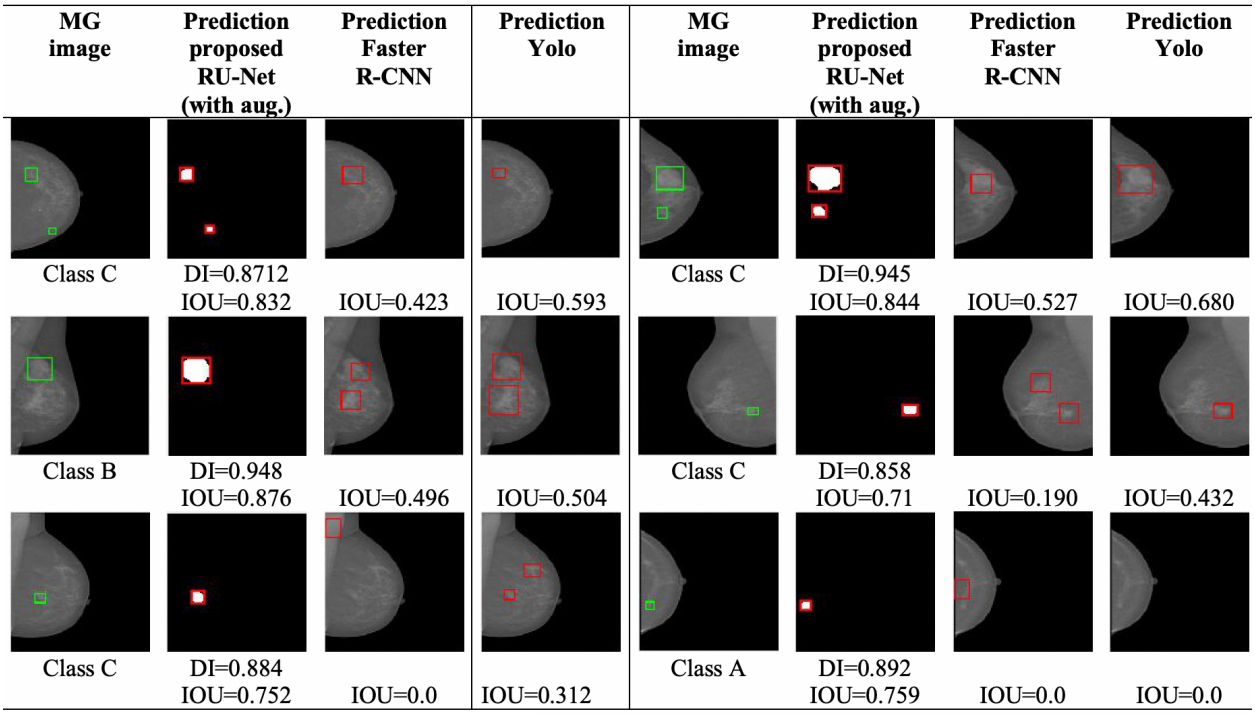
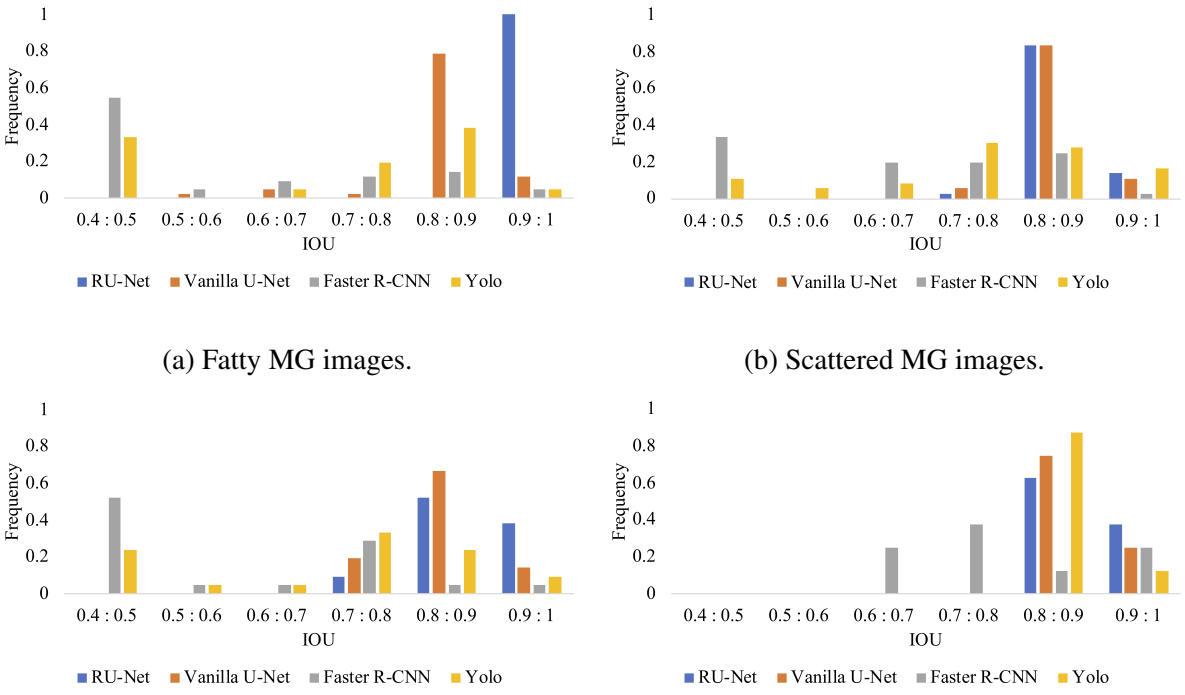


Fig. 9.7: A comparison between the proposed RAttU-Net and the Faster R-CNN and Yolo models in terms of IOU.

values against the GTMs in comparison to the IOU of the Faster R-CNN and the Yolo models, Fig. 9.5 (h and i). However, Yolo provides better precise detection than the Faster R-CNN model due to having adaptive anchor boxes generated from the training data-set.

Detecting small lesions in MGs are very challenging, especially if these small lesions exist in heterogeneously dense and dense MG images. To address the problem of detecting small lesions, the proposed RAttU-Net model uses residual attention blocks to increase the resolution for better pixel-level prediction. By using residual blocks, the network incorporates multi-scale spatial context and captures more local and global context to predict a precise pixel-wise segmentation map of an input full MG image he high-resolution features from the encoder are aggregated more for obtaining stronger semantic information (Fig. 9.5 (e: g)). To obtain more accurate segmentation



(a) Fatty MG images.

(b) Scattered MG images.

(c) Heterogeneous dense MG images.

(d) Dense MG images.

Fig. 9.8: Histogram of the mean of IOU value for the test images in each MG breast density class.

results, we used large images of 640×640 for training the network to provide much contextual information. One of Yolo and Faster R-CNN known drawback is having low detection accuracy on small objects. In Figs. 9.5, 9.6 and 9.7, small lesions in MGs tend to have very low IOU compared to our proposed RAttU-Net.

The proposed RAttU-Net model overcomes the limitation of the state-of-the-art DL segmentation models in terms of reserving high-resolution details by using the residual attention modules, which help the model to detect masses in dense images. Figure 9.6 illustrates the capability of the proposed RAttU-Net in detecting small masses in heterogeneously dense and dense MG images. Moreover, masses that existed over the pectoral muscle in dense areas are detected by DI ≥0.65 (Fig. 9.6, first row). The proposed RAttU-Net model succeeds to detect multiple lesions in the

same breast as shown in Figs. 9.5, 9.6 and 9.7. Moreover, the proposed RAttU-Net can precisely detect multiple lesions with higher DI and IOU in comparison to the other methods (Fig. 9.7, first row).

The Faster R-CNN and the Yolo models can detect lesions in the MG images, however, these models introduce more FPs (In Fig. 9.7, second row). The Faster R-CNN and the Yolo models also results in more false negatives (FNs), as shown in the last row of Fig. 9.7, where they provide very low IOU that reaches 0.0 with FPs, while the proposed RAttU-Net have IOU that ≥ 0.7. As we mentioned before, the MG images of class C (heterogeneously dense) are very challenging where the dense areas of the breast make it harder to find masses and obscure small masses. The proposed RAttU-Net succeed to overcome these challenges (Figs. 9.5, 9.6 and 9.7). This is because the segmentation process incorporates more multi-scale spatial context and captures more local and global context to predict a precise pixel-wise segmentation map of an input full MG image.

We divided the MG images in the INbreast data-set into 4 classes accordingly to its breast density classes. The numbers of images in classes A, B, C, D are 42, 36, 21, 8 images, respectively. We tested the 5 models individually on MG images in each class. Figure 9.8 shows the histogram of the IOU for each class, with a bin width of 0.1. The RAttU-Net detected 100% fatty MG images with an IOU of a range of 1: 0.9 (Fig. 9.8a). The vanilla U-Net follows the proposed network in the detection of masses with 11.9% and 78.57% of fatty MG images having an IOU in the range of 1: 0.9 and 0.9: 0.8, respectively, while the Faster R-CNN and Yolo models show lower IOU in the range of 1: 0.8 (Fig. 9.8b).

In the case of the scattered MG images, the RAttU-Net and the vanilla U-Net have nearly the same IOU histogram distributions in the ranges 1: 0.8 (Fig. 9.8b). In the challenging cases, the hetero-

geneously dense and dense MG images, the proposed RAttU-Net is superior in detecting masses with high IOU in the range of 1: 0.8 than other models (Figs. 9.8c and 9.8d). We noticed that the Faster R-CNN has 43% of all density classes in the range of 0.5: 0.4. The RAttU-Net detected 53.27%, 42.9%, 2.8% of masses in MG images in all density classes with IOU in the range of 1: 0.9, 0.9: 0.8, 0.8: 0.7, respectively. While the vanilla U-Net detected 13.08%, 77.57%, 6.5% of masses in MG images in all density classes with IOU in the range of 1: 0.9, 0.9: 0.8, 0.8: 0.7, respectively (Fig. 9.8). We noticed that the masses that exist in the dense category in the INbreast data-set are relatively larger in terms of size than the other masses in the other categories. That is the reason that why the Faster R-CNN and Yolo models have high IOU in that category (class D) than other categories (classes A, B, C) (Fig. 9.8d).

In Table 9.3, the performance of the models under study is shown for comparison between the tight detected BBs and the ground truth BBs. Moreover, Table 9.3 shows a breakdown of the values of the mean Acc. and mean IOU among different breast density classes. The proposed RAttU-Net is superior in detecting masses than other models understudy with high mean Acc. and mean IOU. The BF-score of the proposed RAttU-Net method is 0.981 which exceeds the other segmentation models under study. The values of DI and IOU of the vanilla U-Net method is closer to that of the RAttU-Net method compared to the original U-Net. The trained original U-Net has a mean DI of 0.756, a mean IOU of 0.836, respectively.

9.5.1 Effect of Augmentation

We investigated the effect of augmentation in the performance of the proposed RAttU-Net method (Fig. 9.5 (e: g)). For example, the values of the DI of the augmented model, as shown

in (f and g), are higher than the ones of the trained model without augmentation (Fig. 9.5 (e)). The mean DI and the IOU of the proposed RAttU-Net with augmentation are 0.983 and 0.948, respectively, compared to 0.905 and 0.891 of the RAttU-Net model without data-augmentation (Table 9.4). The BF-score improves from 0.919 to 0.981 in the case of the proposed augmented U-Net model. Figure 9.5 (e: g), shows that the DI per image increases when the proposed model is trained with the augmented mixed data-set.

9.5.2 Improvements of the Proposed RAttU-Net Model

The proposed model yields an improvement of 4.24%, 30.02% in the DI and 2.93%, 13.39% in the IOU, respectively, relative to that of the vanilla U-Net model and the original U-Net (Table 9.4). The original U-net architecture uses a long skip connection to concatenate the features maps. By replacing the concatenation module with an additional module, the RAttU-Net becomes a fully residual attention model.

9.5.3 Timing Performance

To assess the runtime performance of these methods, we measured the inference time per image taken by each method to detect lesions in the test data-set. The mean inference time per image of the proposed RAttU-Net method, Vanilla U-Net, Original U-Net, Faster R-CNN, and Yolo models are 0.094, 0.087, 0.080, 0.439, 0.206 seconds, respectively. Its running time is comparable (slower in fractions of milliseconds) with the Vanilla U-Net and original U-Net while outperforming the Faster R-CNN, and Yolo models. We have to emphasize that for radiologists, an inference time of

a fraction of second or even several seconds is not as important as the accuracy of the given model.

9.6 Classification Results

We evaluated the performance of the ResNet CNN model in terms of area under the ROC curve (AUC) for the task of classifying the segmented binary maps of breast masses as benign or malignant. Our approach validates the usefulness of using binary GTMs for training the ResNet model. The classification performance is evaluated in terms of sensitivity (Sen.), specificity (Spe.), accuracy (Acc.), F1-score, and Matthews Correlation Coefficient (MCC) per image. Tables 9.5 and 9.6 show the classification performance for the ResNet models trained on augmented binary-scale MG maps and augmented gray-scale MG images using 5-fold cross-validation. Table 9.6 shows the confusion matrix of the classification task. The ResNet model results in a TPR/FPR of 0.94/0.03

Table 9.5: Comparison of the classification performance of the ResNet model over 5-fold cross-validation using data-set trained on augmented binary GTMs images or gray-scale MG images.

Model	Sen.	Spe.	Acc.	AUC	F1-score	MCC
-Proposed ResNet, binary-scale images	0.94	0.96	0.95	0.98	0.93	0.90
-ResNet, gray-scale images	0.87	0.92	0.91	0.96	0.86	0.79

when trained on binary-scale GTMs and a TPR/FPR of 0.87/0.07 when trained on gray-scale images using the INbreast test data-set (Table 9.6). The accuracy of the ResNet model is 0.95 and 0.91 when trained with binary-scale GTMs images and gray-scale images, respectively. It is observed that the ResNet model that is fine-tuned with gray-scale GTMs performs better than the

one fine-tuned with gray-scale MG with mean Sen. (0.94), Spe. (0.96), Acc. (0.95), AUC (0.98), F1-score (0.93), and MCC of (0.90) (Table 9.5). In table 9.6, 94.84% of benign cases and 96.08% of malignant cases are correctly classified, while 3.92% and 5.16% are falsely classified using the ResNet model trained on GTMs (Table 9.6). However, the ResNet model that is fine-tuned using gray-scale images results in 7.63% and 12.58% of miss classified benign cases and malignant cases, respectively (Table 9.6). The results of the classification show the robustness of the proposed

Table 9.6: Confusion matrix of the classification task via the ResNet model trained on binary-scale or gray-scale MG images over 5-fold cross validation.

Model	**Actual classes**	**Predicted classes**	
		Benign	**Malignant**
-Proposed ResNet, binary-scale images	**Benign**	0.94	0.05
	Malignant	0.03	0.96
-ResNet, gray-scale images	**Benign**	0.87	0.12
	Malignant	0.07	0.92

ResNet CNN in minimizing the FP and FN rates (Tables 9.5 and 9.6). The improved performance of the ResNet classifier of binary images is due to the following reasons. First, training GTMs have clear boundaries and margins. Second, the high deep level of features from proposed RAttU-Net contributed to improving the performance of the cascaded classifier by producing precise segmentation maps that can be correctly classified as benign or malignant with high accuracy (Tables 9.5

and 9.6). In our proposed cascaded Res-Net classifier, the model classifies the full detected map. The mean inference time of the classification model is 0.033 sec per image.

9.7 Discussion and Conclusions

We propose a novel network architecture, which consists of two cascaded CNNs. The first network is a residual U-Net (RAttU-Net) for semantic segmentation of mass lesions in MG images. The proposed RAttU-Net predicts a pixel-wise segmentation binary map of an input full MG image efficiently due to the residual attention modules. The second network, a ResNet classifier, is used for the task of binary shape classification of the segmented binary-scale maps into benign and malignant. We compared the performance of the proposed RAttU-Net model with the performance of two DL semantic segmentation models, U-Net and Vanilla U-Net, and two DL object detector models, YOLO and Faster R-CNN. We trained all the models with the same data-sets. We observed that the data augmentation used to increase the training data-set size enhances the performance of the proposed model. We also observed that the proposed model has a lower mean inference run-time per image compared to the other DL models. The proposed RAttU-Net model achieves a mean test pixel Acc.of 0.98, mean DI of 0.98 and mean IOU of 0.94 that outperform those of the other models. In summary, the proposed RAttU-Net model can be used for the precise segmentation of masses in MG images, especially for the challenging heterogeneously dense and dense breast cases.

The proposed RAttU-Net is superior in detecting masses in dense MGs with high IOU in the range of 0.8: 1 than other models. This is because the segmentation process incorporates a more multi-scale spatial context and captures a more local and global context to predict a precise pixel-

wise segmentation map of an input full MG image. The results show that the precise segmented masses can be used for more accurately differentiating benign from malignant lesions that is a very challenging task. The fine-tuned ResNet model with binary-scale GTMs performs better than the ResNet model fine-tuned using gray-scale MG images in terms of mean Sen. (0.94), Spe. (0.96), Acc. (0.95), AUC (0.98), F1-score (0.93), and MCC (0.90). To conclude, using TL, introducing Aug., and incorporating multi-scale local and global context using the residual attention modules into the original U-Net architecture result in a better performance in detecting and segmenting masses which can be used more effective classification of masses.

Chapter 10

Density Attention-based Convolution Neural Network for Breast Cancer Detection in Mammogram Images.

10.1 Abstract

The detection of cancer in dense tissues and, more generally, in dense breasts is often considered more challenging due to the similar visual aspects of normal and abnormal tissues, which complicates the interpretation of mammographic images. The density of MGs makes mass detection very challenging since masses can be hidden in dense mammograms (MGs). For the above reasons, we argue that deep learning (DL) used for lesion detection should first evaluate the spatial distribution of the dense tissues. In this research, we first generate an attention breast density mask, identifying pixels associated with the tissue that contributed to the density class. These generated attention masks to identify and exploit the effective dense segments of MGs images to support the CNN detection decision. The approach is premised on the hypnosis that there is benefit from amplifying the influence of dense regions, compared to treat a whole MG the same, especially for the heterogeneously dense and scattered regions. Attention masks of breast density will be extracted automatically using a residual attention U-Net model. To target the above challenges, the generated attention maps will help the main CNN used for detecting lesions to pay attention more

to these regions. We compared the performance of the nested U-Net (U-Net++) model, vanilla U-Net, recurrent residual U-Net (R2U-Net), attention U-Net (AttU-Net), residual attention U-Net (RAttU-Net) that are trained from scratch on the segmentation of lesions with the corresponding models that are trained using the proposed training strategy. The RAttU-Net is superior in detecting masses in heterogeneous and dense MGs with high mean intersection over union (IOU) of 0.912 and 0.858, respectively than other models. Following it is the AttU-Net (pre-trained) with a high IOU of 0.858 and 0.833 in detecting masses in heterogeneous and dense MG, respectively.

10.2 Background

Breast cancer is the second leading cause of cancer death in women. In 2019, an estimated 268,600 new cases of invasive breast cancer are expected to be diagnosed in women and an estimated 41,760 women with breast cancer are expected to die in the USA [204]. From 2007 to 2016, the breast cancer death rate declined by 1.8% per year. This decline in breast cancer mortality has been attributed to improvements in early detection through early screening using mammograms (MGs) and treatment, as well as advances in the detection techniques using computer-aided detection (CAD) models [204].

Mammographic density (MD) is a strong breast cancer risk factor. The Breast Imaging Reporting and Database Systems, or BI-RADS, which reports the findings of mammograms, also includes an assessment of MD. BI-RADS classifies MD into four classes: BI-RADS (A): almost entirely fatty (0-25%), BI-RADS (B): scattered with some fibroglandular tissue (26-50%), BI-RADS (C): heterogeneously dense (51-75%), and BI-RADS (D): extremely dense (76-100%) [10, 11], as in Table 10.1. Researchers combine the classes of heterogeneously dense and extremely dense and

categorize them as dense areas (DA), whereas the patterns of fatty and scattered density are labeled as non-dense areas (NDA) [4, 10, 204].

A study in [206] states that about 43% USA women aged 40 to 74 years have heterogeneously or extremely dense breasts. Mammographic detection of breast cancer is impaired in areas of dense breast tissue [11]. Studies have shown that the sensitivity (Sen.) of CAD models using MGs significantly decrease as the density of the breast increases [10,11,96,126,128,152]. The sensitivity of mammography for women with almost entirely fatty breasts is 88% as compared with 69% for women with heterogeneously dense breasts, and 62% for women with extremely dense breasts [227]. Compared to women with NDA, those with DA have about a 1.6 or 2.3 times, respectively, higher risk of breast cancer [204]. Several studies have classified the MD according to BI-RADS classes [128], based on features extracted from the MGs using texture or gray-level information, as in Table 10.1. In respect of the breast density classification, Mohamed et al. [152] investigated

Table 10.1: Distribution of tissue density according to BI-RADS specification.

BI-RADS	Density	Breast density	Remarks
A	0 – 25%	Entirely fatty tissue.	-
B	26 – 50%	Scattered fibroglandular tissue.	-
C	51 – 75%	Heterogeneously dense tissue.	Obscure small masses.
D	76 – 100%	Extremely dense tissue.	Lowers sens. of MGs.

a DL-based breast density classifier based on the AlexNet model pre-trained on ImageNet and fine-tuned it with mammogram (MG) data to consistently distinguish between scattered dense and heterogeneously dense MG images. Mohamed et al. show that using the transfer learning (TL)

method from non-medical images to medical images boosted the performance of their CNN-based classifier from AUC of 0.94 to AUC of 0.98. In another study, Ahn et al [15] presented a CNN-based approach for breast density estimation. The CNN was trained to learn image features from the patches extracted from the whole MGs and classify them as fatty and dense class tissues. Lehman et al. [134] trained a ResNet-18 model to classify MG images into new categories dense and non-dense. Lehman et al. trained his model to classify BI-RADS categories C and D as dense MG images and BI-RADS categories A and B as non-dense MG images. In a similar study to [134], Xu et al. [241] classified the breast density MG images using residual-based CNN. Their proposed method obtained an overall accuracy of 92.63% for the four BI-RADS class classification task and accuracy of 96.84% for the two BI-RADS class classification task (i.e. scattered fibroglandular and heterogeneous dense). The authors used BI-RADS A and B as one scattered density class and BI-RADS C and D as one heterogeneously dense class to train their model. Similarly, Wu et al [239] presented the application of deep neural network (DNN) for the classification of breast densities in digital MGs. The study comprised 20,000 screening mammograms labeled as 4 class breast densities. Additionally, they also considered accuracy only between the two super-classes: class C and D (dense) versus class A and B (not dense). A scratch-based CNN with dense convolutional layers was used to discriminate the breast densities in the multi-view data.

Kallenberg et al [115] proposed an unsupervised DL technique to segment dense regions of the breast density. The method uses a conventional sparse autoencoder (CSAE) for learning the features. Geras et al [79] in their study used DL CNN for the prediction of breast densities in multi-view data. The method predicted breast density and classified into 3 types: BI-RADS 0, BI-RADS 1, and BI-RADS 2. In this work, we developed a deep learning model that analyzes

an MG image to find dense areas (DA) that may have a higher risk of masking cancer than other non-dense areas (NDA).

10.3 Materials and Methods

10.3.1 Data-sets

We trained our CNN models on two the publicly available data-sets, DDSM [93] and BCDR [142], and tested them on the publicly available INbreast data-set [38] and the private UCHCDM data-set. UCHCDM is a fully digital data-set collected at the University of Connecticut Health Center (UCHC) [9, 253]. Each MG image in UCHCDM has a large size (e.g., 3328×4096 pixels) and 12 bits per pixel. UCHCDM images are originally stored in DICOM format. All MG images are ground truth annotated files with a description of known pathology (normal (healthy), mass, architectural distortion (AD), calcification (calc)) and circled at the locations of cancers (if any) on a separate key ground truth image.

To have ground truth for evaluating state-of-the-art object detection methods, we generated ground truth bounding boxes (BBs) for the masses based on minimum and maximum points values of x and y coordinates of the mass' contours, which indicate the locations of masses. Each MG image has been annotated based on their density derived from the American College of Radiology's (ACR) Breast Imaging Reporting and Data System (BI-RADS) [29]. For each MG image, its density in ACR standard scale is given as one of these categories: class A: fatty, class B: scattered, class C: heterogeneous dense, and class D: dense [4, 29]. We grouped the BI-RADS multi-class assessment into benign and malignant classes. In this study, we categorized 1133 MG images with BI-RADS $\in$ {2, 3} as benign, and 2093 MG images with BI-RADS $\in$ {4, 5, 6} as malignant. The

distribution of density for each BI-RADS class is presented in Table 10.2. In total, we used 3,226 MG images to conduct our experiments.

Table 10.2: Distribution of breast density in each BI-RADS class in the data-sets used in our study.

Purpose	Data-set	BI-RADS	Class				Total
			A	B	C	D	All
Training	**DDSM**	Benign	54	215	228	224	1,018
Training	**BCDR-D01**	Benign	34	16	23	7	80
Test	**INbreast**	Benign	12	4	13	6	35
Training	**DDSM**	Malignant	219	693	513	291	1,716
Training	**BCDR-D01**	Malignant	16	15	22	2	55
Test	**INbreast**	Malignant	30	32	8	2	72
Test	**UCHCDM**	Malignant	20	164	66	0	250

10.3.2 Data Pre-Processing

The mammographic images are first pre-processed by employing normalization by scaling the image intensity to the range of [0, 255]. To pre-train our models for segmentation of dense areas in MG images, we first detect the breast boundary then all pixels exists in the black background areas takes values 0, all breast non dense areas takes values (128), and dense areas takes values (255) classes. After several experiments on different MG data-sets, we find the dense areas lies between the intensities 130 and 255, thus we used these pixels values as the ground truth for dense areas

within the breast. No augmentation technique was used to train our models to segment dense areas within the breast.

To segment mass(es) in the MG images, we generated different ground truth maps (GTMs) for the masses using the associated pixel-level boundary of the mass lesions given by the data-sets. All pixels in the GTM are labeled as belonging to the background (0) or breast lesion (255) classes. To deal with the small training data-set and avoiding overfitting our model, we applied data augmentation to the training MG images and it's corresponding GTMs (that contain mass(es) by image rotation by (-15, 15) degrees, translation up and down by (-10%, 10%), scaling in and out by 0.2, and left-right flips. All full MGs images and it's corresponding GTMs are re-sized to 512×512. The images that belong to the same patient are either in the training data-set or in the test data-set.

10.3.3 Proposed Training Approach

The proposed approach is premised on the hypnosis that there is benefit from amplifying the influence of dense regions, compared to treat a whole MG the same, especially for the heterogeneously dense and scattered regions. These dense regions have a similar texture to the mass lesions. The idea is to give these regions higher weight than other regions in the breast while the network is being trained to detect lesions. To handle this problem, we propose to pre-train the DL segmentation models first to detect dense regions in MG images. This will make the network learn different features for this region than other non-dense regions in the MG image. Thus for every MG image, the network will generate attention maps of breast density. The generated attention maps will help the main CNN used for detecting lesions to pay attention more to these regions. Next, the pre-trained network is fine-tuned with augmented ground truth masks (GTMs) that show

the location of the tumors. We did not use any augmentation technique for pre-training our models. However, we used data augmentation while we fine-tune the models on GTMs of lesions.

10.3.4 Training Models

In this work, we apply our training technique on different segmentation models and compare its performance in segmentation of mass lesions before and after applying it. We compared the performance of the nested U-Net (U-Net++) model [255], Vanilla U-Net [6, 10], recurrent residual U-Net (R2U-Net) [23], attention U-Net (AttU-Net) [159], residual attention U-Net (RAttU-Net) [7] that are trained from scratch on the segmentation of lesions with the corresponding models that are pre-trained on segmenting dense areas. U-Net is a popular E2E encoder-decoder model for semantic segmentation that is originally invented for bio-medical image segmentation [186]. U-Net consists of a contracting path to capture features and an asymmetric expanding path that enables precise localization and segmentation of pixels. This architecture has a U-shaped skipping structure that connects the high-resolution features from the contracting path to the up-sampled outputs of expanding path.

The vanilla U-Net [10] provides better architecture than the original U-Net by using batch normalization (BN) layers, dropout layers of (0.5) and the use of convolution blocks instead of the regular convolution layers (see Fig. 10.1a).

In the UNet++ architecture is a deeply-supervised encoder-decoder model where the encoder and decoder sub-networks are connected through a series of nested, dense skip pathways. The re-designed skip pathways aim at reducing the semantic gap between the feature maps of the encoder and decoder sub-networks.

Recurrent Residual Convolutional Neural Network (R2U-Net) [23] utilizes the power of U-Net,

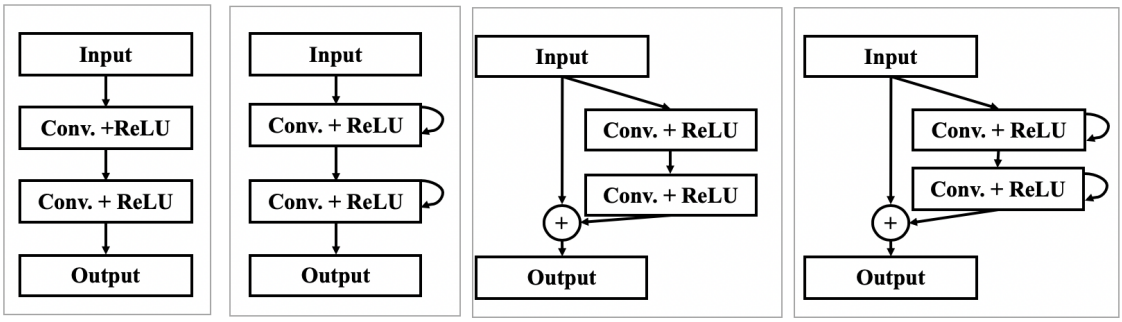
Residual Network, as well as recurrent CNN. The advantage of this architecture is that the residual unit helps when training deep architecture. Moreover, the feature accumulation with recurrent residual convolutional layers ensures better feature representation for segmentation tasks. It also has the same number of network parameters as U-Net but with better performance for medical image segmentation (see Fig. 10.1d).

In attention U-Net (AttU-Net) [159], the attention gates (AGs) are incorporated into the standard U-Net architecture to highlight salient features that are passed through the skip connections, (see Fig. 10.1e). Information extracted from coarse-scale is used in gating to disambiguate irrelevant and noisy responses in skip connections. This is performed right before the concatenation operation to merge only relevant activations. Additionally, AGs filter the neuron activations during the forward pass as well as during the backward pass.

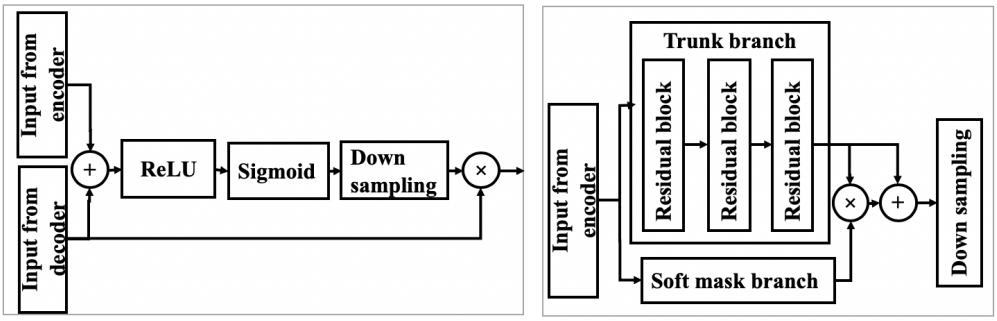
Inspired by the residual attention mechanism proposed in [229], the authors in [7] built the RAttU-Net model by stacking residual attention modules to the basic U-Net architecture. The authors used residual blocks with identity connections instead of the regular convolution layers in the U-Net architecture to preserve the spatial and context information, help the network have deeper architecture, and handles the gradient vanishing problem. The residual blocks directly propagate features from its early convolution to its late convolution and improve the performance of the model consequently. To address the problem of detecting small lesions, the RAttU-Net model uses residual attention blocks to increase the resolution for better pixel-level prediction (Fig. 10.1f).

10.3.5 Training Configurations

For training the vanilla U-Net, AttU-Net, RAttU-Net segmentation models, we used the Dice co-



(a) Forward conv. block. **(b) Recurrent conv. block.** **(c) Residual conv. block.** **(d) Recurrent residual conv. block.**



(e) Attention gate. **(f) Residual attention gate.**

Fig. 10.1: Building blocks of neural networks. (a) forward convolutional units, (b) recurrent convolutional block (c) residual convolutional unit, (d) recurrent residual convolutional units (RRCU), (e) attention gate, and (f) residual attention gate.

efficient (DC) loss as the objective function to train the model. The DC loss function is minimized using Adam optimizer with a decreasing learning rate (LR) initialized to 10^{-3} and a momentum of 0.9. The LR is reduced every 25 epochs by a factor of 0.1. To train the R2U-Net, we used Adam optimizer with an LR of 2×10^{-4} and binary cross-entropy loss. To train the UNet++ model, we used Adam optimizer with an LR of 3×10^{-4}. The loss function for training the UNet++ model is a combination of binary cross-entropy and dice coefficient [255]. We trained all the models for 200 epochs using mini-batches of size 2. To manage imbalanced data, we introduced class weights into the loss functions.

10.3.6 Evaluation Metrics

To evaluate the performance of the DL models, the Dice index (DI), also known as the F1 score, and the IOU, also known as the Jaccard index, metrics are used to compare the automated predicted maps with the GTMs. We considered a detected segment as true positive (TP) if it overlaps with the ground truth segment by more than 40%.

10.4 Experiments and Results

To demonstrate the effects of our training technique, we performed experiments on two different medical mammogram data-sets (INbreast and UCHCDM) which include 2D images for breast lesion segmentation. We use the PyTorch framework to implement all the experiments on a single GPU machine on Tesla V100-PCIE-16GB.

10.4.1 Segmentation Results for Dense Regions

After pre-train our models on background, dense, and non dense area classes, we set a threshold for the segmented dense tissue; the area of dense tissue is expressed in one of four-category

percentages: 0 – 25%, 26 – 50%, 51 – 75%, and 76 – 100% in reference to the BI-RADS levels. According to the previous defined criteria, of the 62 fatty MG images there are 60 images whose breast percentage density (PD) between 10 – 25% (accuracy of 96.774%). Out of the 200 scattered MG images there are 12 images with PD less than %26 and ten MG images with PD greater 50% and misclassified as heterogeneous class (accuracy of 90%). Out of the 102 heterogeneous MG images there were 90 images with PD in range 51 – 75%, 10 MG images misclassified as heterogeneous class and 2 MG images misclassified as dense class (accuracy of 88.235%). ALL the 8 dense images were correctly classified as dense class (accuracy of 100%). Table 10.3 shows the confusion matrix of the test data-sets in the four BI-RAD classes. The results show that the segmentation output show similarity in some of the images between the scattered and heterogeneous classes. Figure 10.2 shows MG images in BI-RAD A, B, C, and D and their corresponding binary segmented output maps using the RAttU-Net model to segment it to background (black areas), dense (white areas) and non dense (gray) areas. The total accuracy of the classification task to the four BI-RAD classes is 0.91 and AUC of 0.852. The benefit of this pre-training is to make the networks have different wights for dense regions in MG images.

10.4.2 Segmentation Results for Mass Lesions

We compared the performance of the nested U-Net (U-Net++) model [255], vanilla U-Net [6], recurrent residual U-Net (R2U-Net) [23], attention U-Net (AttU-Net) [159], residual attention U-Net (RAttU-Net) [7] that are trained from scratch on the segmentation of lesions with the corresponding models that are pre-trained on segmenting dense areas. Furthermore, we trained the state-of-the-arts, the Faster R-CNN model proposed in [185], and the Yolo model proposed in [21] with our training data-sets from scratch to detect masses. We use the mean DI, mean IOU, and the infer-

Table 10.3: The confusion matrix of the classification of the INbreast and UCHCDM into the four BI-RAD classes.

		Truth data			
		Class A	**Class B**	**Class C**	**Class D**
Classifier results	**Class A**	60	12	0	0
	Class B	2	180	10	0
	Class C	8	8	90	0
	Class D	0	0	2	8
	Accuracy	96.774%	90.0%	88.235%	100%
	Precision	83.333%	93.75%	91.837%	80%

ence time in seconds per image to measure the performance of each model. Further details about the implementation of these models can be found in the original article in [6,7,21,23,159,185,255]. We trained the above models using our data-sets and tested them using the INbreast and UCHCDM data-sets.

Figures 10.5, 10.6, 10.7, 10.8, 10.9, 10.10, 10.13, 10.14, 10.15, 10.16, 10.17, 10.11, 10.12 , 10.18, 10.19, **??** and 10.20 (a) show the original FFDM MG images from the INbreast and UCHCDM data-sets, where the red BBs show the location of lesions as given by experienced radiologists, (b) shows the associated pixel-level GTMs of the mass lesions, (c) and (d) the output of U-Net++ with/without pre-training on dense areas, (e) and (f) the output of vanilla U-Net with/without pre-training on dense areas, (g) and (h) the output of R2U-Net with/without pre-

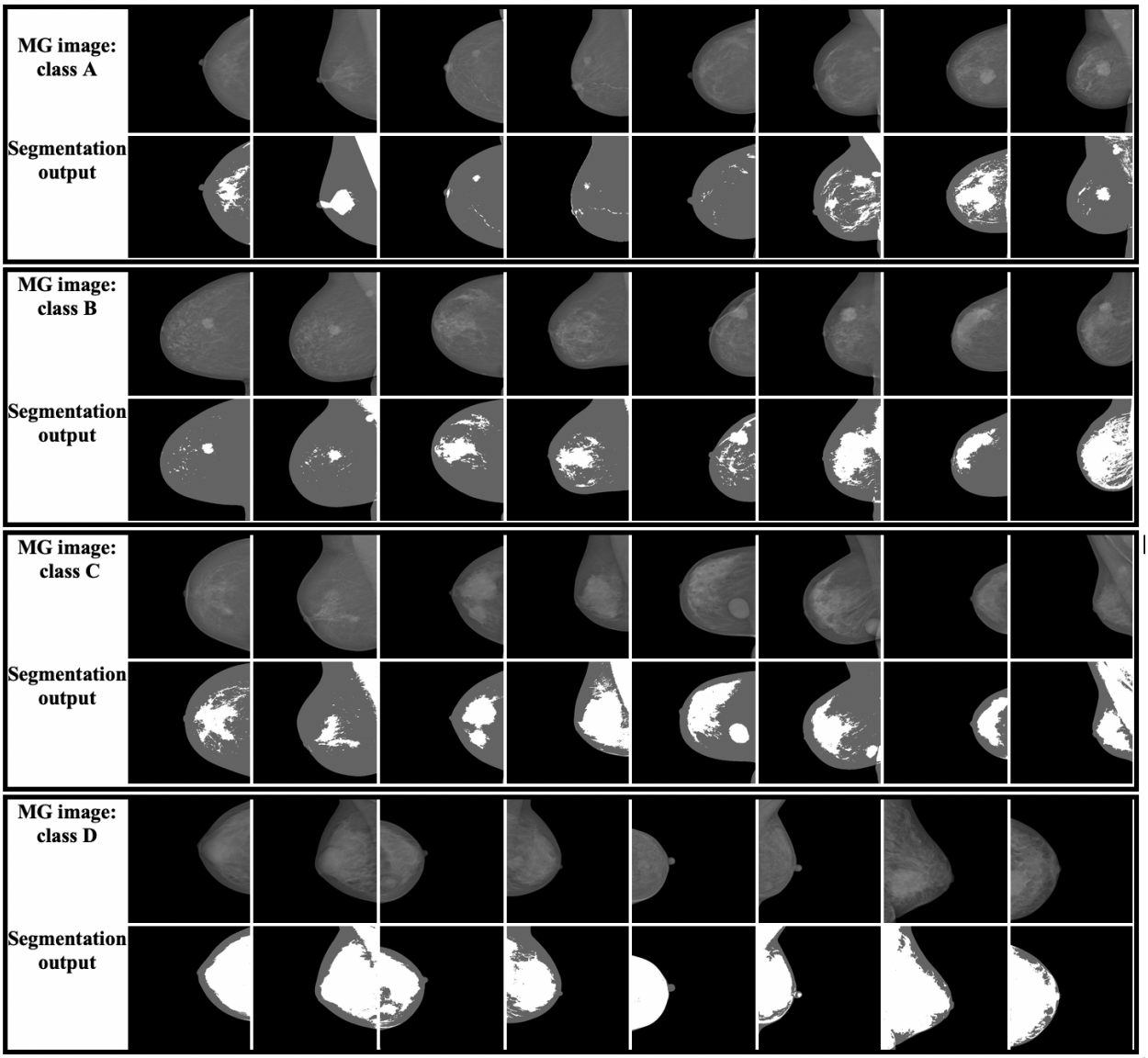


Fig. 10.2: MG images in BI-RAD A, B, C, and D and their corresponding binary segmented output maps using the RAttU-Net model to segment it to background (black areas), dense (white areas) and non dense (gray) areas.

training on dense areas; (i) and (j) the output of AttU-Net with/without pre-training on dense areas and finally (k) and (l) the output of RAttU-Net with/without pre-training on dense areas. Figures 10.5, 10.6, 10.7, 10.8, and 10.9 show the output of the models in class A. Figures 10.13, 10.10,

10.11, 10.12, 10.13, 10.14, 10.15, 10.16, and 10.17 show the output of the models in class B. Figures 10.18, 10.19, and **??** show the output of the models in class C. And finally, Fig. 10.20 show the output of the models in class D. The calculated DI and/or IOU for each prediction is shown under each image in the above mentioned Figures. The DI number shows the Dice similarity coefficient between the current predicted map and its corresponding GTM. DI takes a value in the range [0, 1], where 1 means that the segmentation in the two images is a perfect match. The model that has higher detection in terms of DI than the other models is labeled in bold. We also labeled each MG image with the corresponding breast density class.We added these labels to visually find the effect of breast density on the detection of masses in MGs across different models. Classes A, B, C, and D are corresponding to fatty, scattered, heterogeneously dense, and dense classes, respectively. We divided the MG images in the INbreast and UCHCDM data-set into 4 classes accordingly to its breast density classes. We tested the 5 models individually on MG images in each class.

The performance of the U-Net++, vanilla U-Net, R2Net, AttU-Net, RAttU-Net, Faster R-CNN, and Yolo in terms of the IOU are shown in Table 10.4. The top two models are labeled in bold Table 10.4. The RAttU-Net (pre-trained) is superior in detecting masses in heterogeneous and dense MGs with a high mean IOU of 0.912 and 0.858, respectively than other models. Following it is the AttU-Net (pre-trained) with a high IOU of 0.858 and 0.833, respectively. The RAttU-Net (pre-trained) and U-Net++ (pre-trained) show near mean IOU. The best model show improvement in detection of tumors with IOU increase from 0.923 to 0.950 (fatty class), from 0.887 to 0.928 (scattered class), from 0.805 to 0.912 (heterogeneous class), and 0.757 to 0.858 (dense class). Note that the UCHCDM data-set does not have GT contours for the lesions, however, it has a rectangular bounding box that includes the tumors, which might decrease the DI values of the detected tumors.

Figures 10.3 and 10.4 show the histogram of the IOU for each class, with bin width of 0.2. The RAttU-Net detected 71.43% fatty MG images with an IOU of range of 1: 0.8 (see Fig. 10.3a). The AttU-Net follows the RU-Net network in the detection of masses with 16.67% and 57.14% of fatty MG images having an IOU in the range of 1: 0.8 and 0.8: 0.6, respectively. while the Vanilla U-Net (50.00%), U-NET++ (50.00%), Yolo (42.86%), Faster R-CNN (19.05%), R2U-Net (19.05%) models show lower IOU in the range of 1: 0.8 (see Fig. 10.3a). In the case of the scattered MG images, the RAttU-Net have IOU of 75.00% followed by the U-NET++ (55.56%), AttU-Net (52.78%), Vanilla U-Net (44.44%), Yolo (44.44%), R2U-Net (30.56%) and Faster R-CNN (27.78%) in the ranges 1: 0.8 (see Fig. 10.3d).

In the challenging cases, the heterogeneously dense, the RAttU-Net is superior in detecting masses with IOU (52.38%) in the range of 1: 0.8 than other models (see Fig. 10.3c), followed by the U-NET++ (38.10%), YOLO (33.33%), Vanilla U-Net (28.57%), AttU-Net (28.57%), Faster R-CNN (9.52%), and R2U-Net (4.76%).

In the most challenging cases, the extremely dense, the RAttU-Net has IOU (62.50%) followed by the YOLO (50.00%), AttU-Net (37.50%), Faster R-CNN (37.50%), Vanilla U-Net (25.00%), U-NET++ (25.00%), and R2U-Net (0%) (Fig. 10.3d).

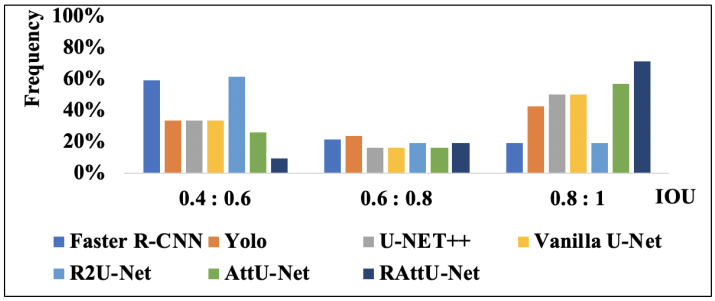
We noticed that RU-Net has the lowest performance in terms of DI and IOU among all the other models and the RAttU-Net have the highest performance in terms of DI and IOU among all the other models. After pre-train the segmentation models on the detection of dense areas in the MG breast, the performance of the models in terms of DI and IOU increased. Figs. 10.3 and 10.4 show that the IOU in the range of 1: 0.8 of the RAttU-Net increased from 71.43% to 88.10% in fatty class, from 75.00% to 75.00% scattered class, from 52.38% to 80.95% in heterogeneous class, and

62.50% to 75.00% in dense class.

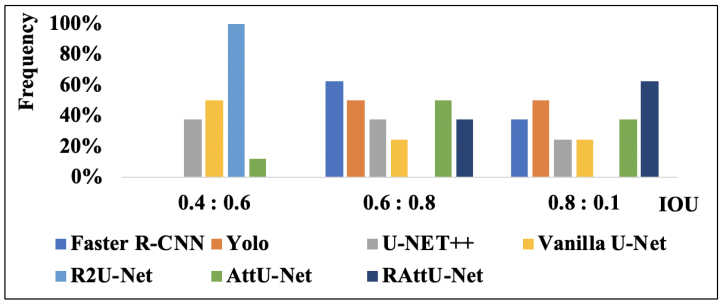
The IOU in the range of 1: 0.8 of the AttU-Net increased from 57.14% to 66.67% in fatty class, from 52.78% to 55.56% scattered class, from 28.57% to 47.62% in heterogeneous class, and 37.50% to 62.50% in dense class (Figs. 10.3 and 10.4). Moreover, the IOU in the range of 1: 0.8 of the U-Net++ increased from 50.00% to 61.90% in fatty class, from 55.56% to 61.11% scattered class, from 38.10% to 47.62% in heterogeneous class, and 25.00% to 37.50% in dense class (Figs. 10.3 and 10.4). The IOU in the range of 1: 0.8 of the Vanilla U-Net increased from 57.14% to 66.67% in fatty class, from 52.78% to 55.56% scattered class, from 28.57% to 47.62% in heterogeneous class, and 37.50% to 62.50% in dense class. The IOU in the range of 1: 0.8 of the U-Net++ increased from 50.00% to 83.33% in fatty class, from 44.44% to 58.33% scattered class, from 28.57% to 57.14% in heterogeneous class, and 25.00% to 37.50% in dense class (Figs. 10.3 and 10.4). And finally the IOU in the range of 1: 0.8 of the R2U-net increased from 19.05% to 83.33% in fatty class, from 30.56% to 47.22% scattered class, from 4.76% to 47.62% in heterogeneous class, and 0.00% to 12.50% in dense class (Figs. 10.3 and 10.4). Even-though the performance of the R2-Net has improved after our training strategy, it still shows the lowest performance among other models in terms of DI and IOU.

Timing performance

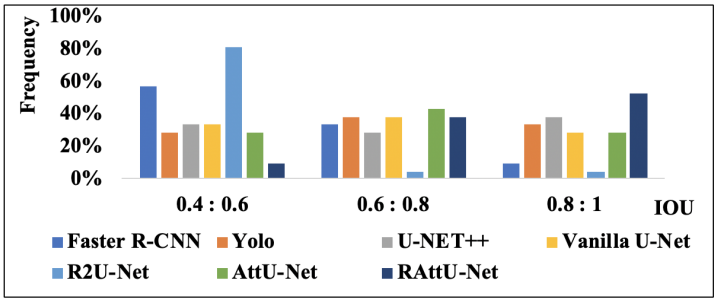
To assess the runtime performance of these methods, we measured the inference time per image taken by each method to detect lesions in the test data-set. We also measured the time taken by each model to complete one epoch. The mean training time per epoch and the mean inference time per image of the previous models are shown in Table 10.5. The inference time of the segmentation



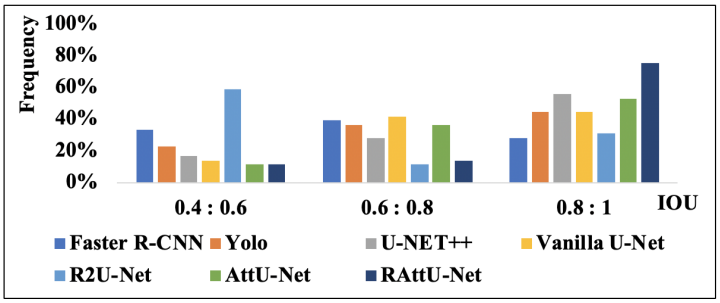
(a) Class A (almost fatty)



(b) Class B (scattered fibroglandular)

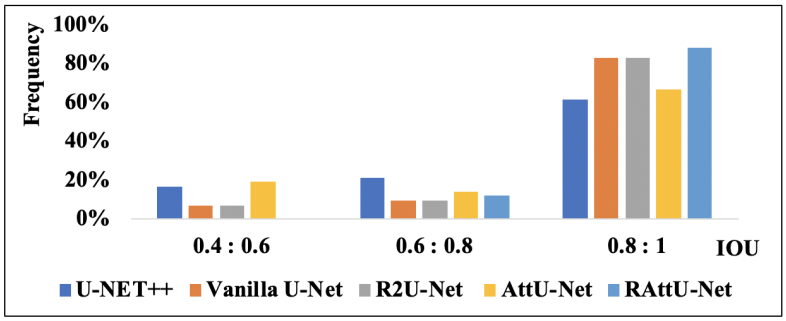


(c) Class C (heterogeneously dense)

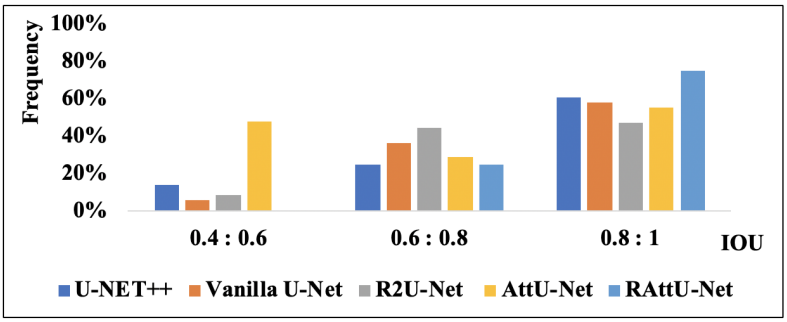


(d) Class D (extremely dense)

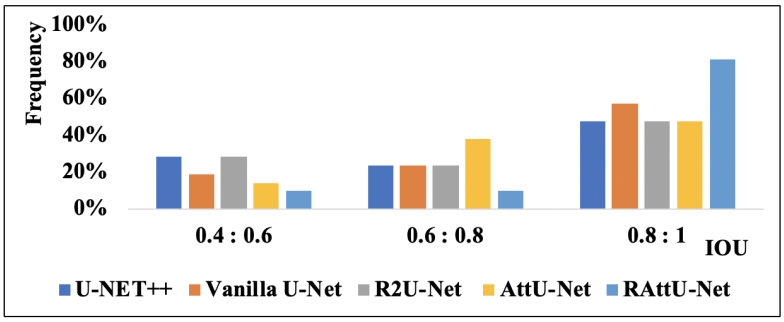
Fig. 10.3: Histogram of the mean of IOU value for the test images in each MG breast density class. Models are trained from scratch on mass GTMs.



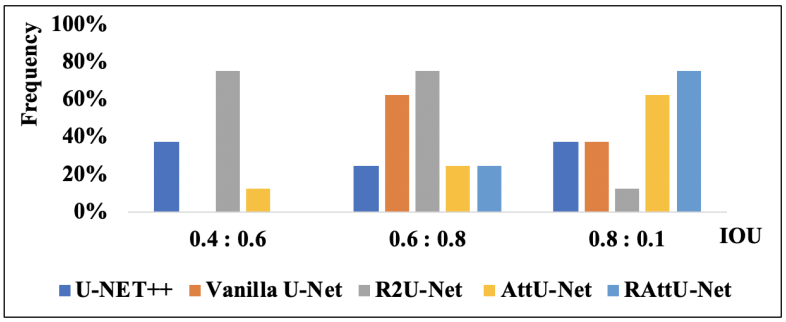
(a) Class A (fatty)



(b) Class B (scattered fibroglandular)



(c) Class C (heterogeneously dense)



(d) Class D (extremely dense)

Fig. 10.4: Histogram of the mean of IOU value for the test images in each MG breast density class. Models are pre-trained with dense images and fine-tuned with mass GTMs.

Table 10.4: The performance of the U-Net++, vanilla U-Net, R2Net, AttU-Net, RAttU-Net, Faster R-CNN, and Yolo in terms of the intersection over union (IOU).

	Class A	**Class B**	**Class C**	**Class D**
Model	**Mean IOU**			
-U-Net++	0.885	0.872	0.710	0.435
-U-Net++ (pre-trained)	**0.942**	**0.915**	0.747	0.533
-Vanilla U-Net	0.930	0.883	0.779	0.601
-Vanilla U-Net (pre-trained)	0.910	0.820	0.807	0.712
-R2U-Net	0.675	0.601	0.476	0.105
-R2U-Net (pre-trained)	0.857	0.718	0.756	0.543
-AttU-Net	0.897	0.834	0.785	0.603
-AttU-Net (pre-trained)	0.943	0.854	**0.858**	**0.833**
-RAttU-Net	0.923	0.887	0.805	0.757
-RAttU-Net (pre-trained)	**0.950**	**0.928**	**0.912**	**0.858**
-Faster R-CNN	0.717	0.732	0.689	0.750
-Yolo	0.814	0.791	0.778	0.802

models are comparable (fractions of milliseconds), while outperforming the Faster R-CNN, and Yolo models.

Table 10.5: The performance of the U-Net++, vanilla U-Net, R2Net, AttU-Net, RAttU-Net, Faster R-CNN, and Yolo in terms of training time per epoch and the mean inference time in seconds.

Model	Mean training time (sec)/epoch	Mean inference time(sec)/epoch
-U-Net++	45	0.013
-Vanilla U-Net	35	0.006
-R2U-Net	37	0.007
-AttU-Net	39	0.011
-RAttU-Net	44	0.012
-Faster R-CNN	65	0.439
-Yolo	70	0.206

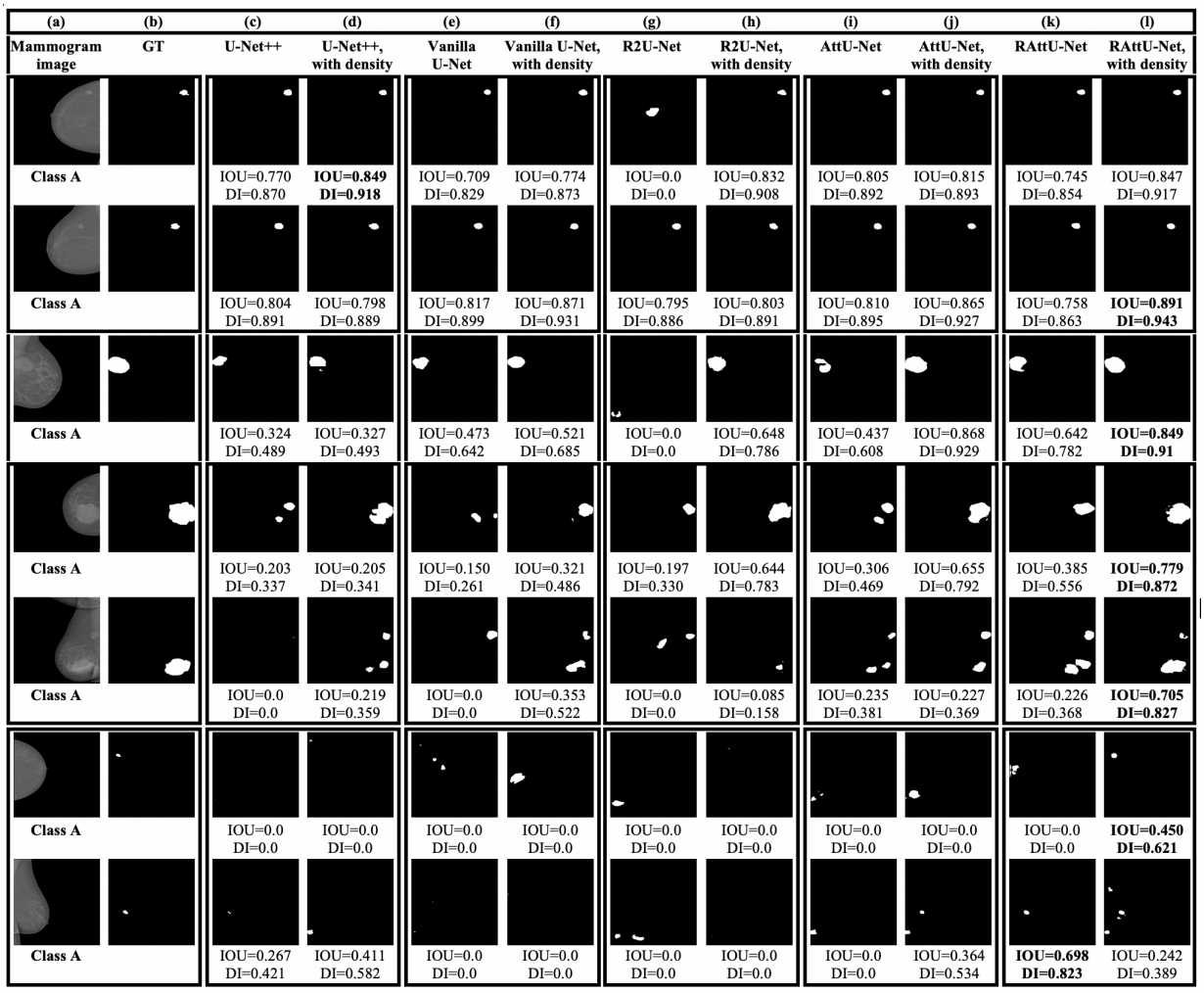


Fig. 10.5: (a) shows class A INbreast MG images and its corresponding GT mask (b); columns (c), (e), (g), (i), and (k) show the output of the trained models without pre-training on dense areas; (d), (f), (j), and (l) show the output of the trained models pre-trained on dense areas.

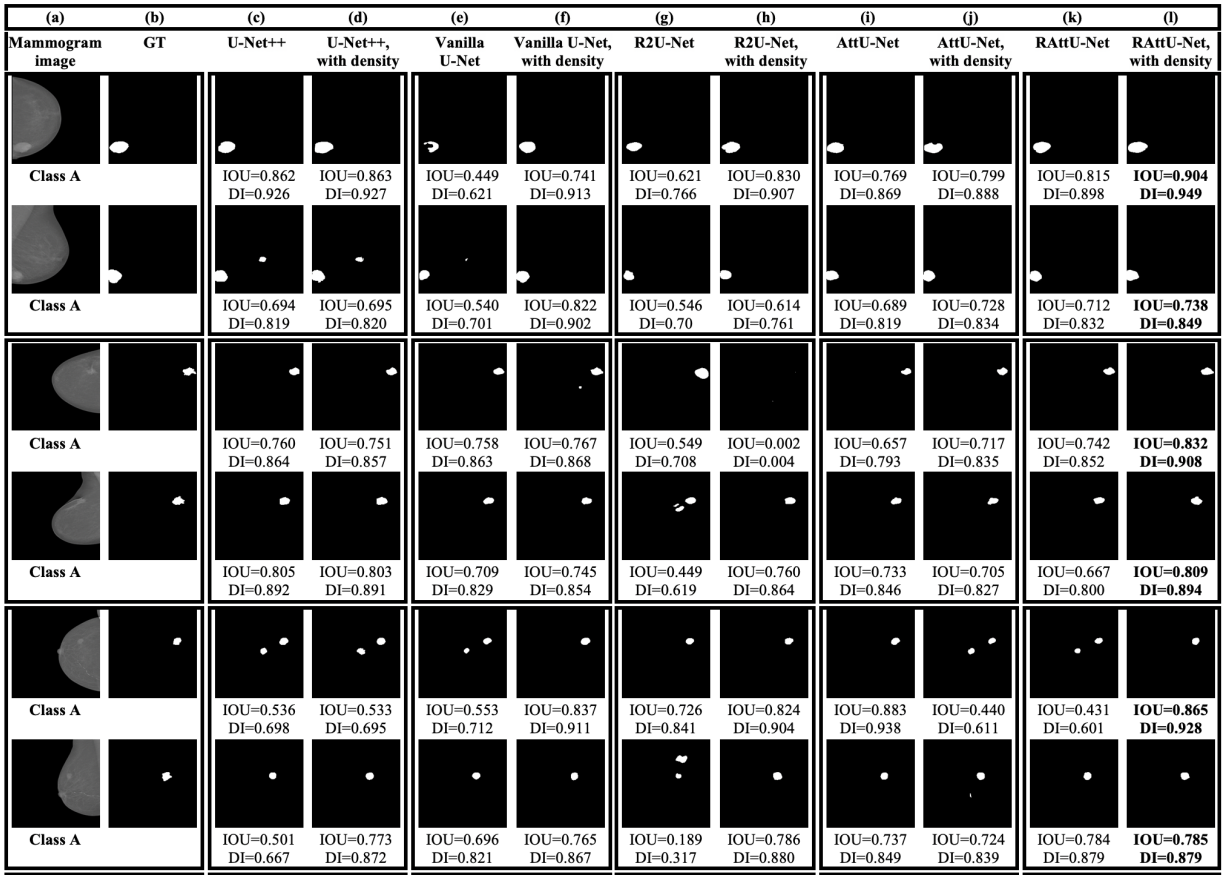


Fig. 10.6: (a) shows class A INbreast MG images and its corresponding GT mask (b); columns (c), (e), (g), (i), and (k) show the output of the trained models without pre-training on dense areas; (d), (f), (j), and (l) show the output of the trained models pre-trained on dense areas.

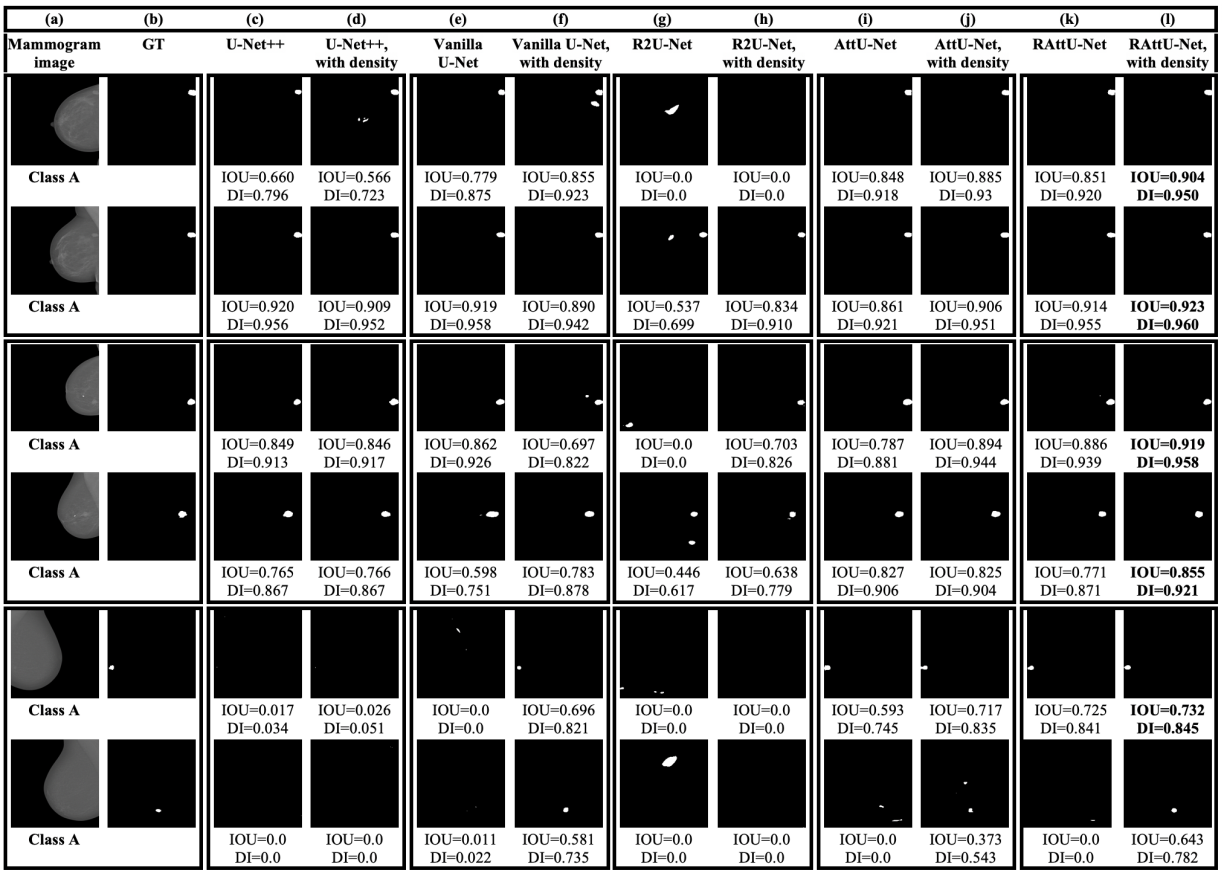


Fig. 10.7: (a) shows class A MG images of the INbreast data-set and its corresponding GT mask in (b); columns (c), (e), (g), (i), and (k) show the output of the trained models without pre-training on dense areas; (d), (f), (j), and (l) show the output of the trained models pre-trained on dense areas.

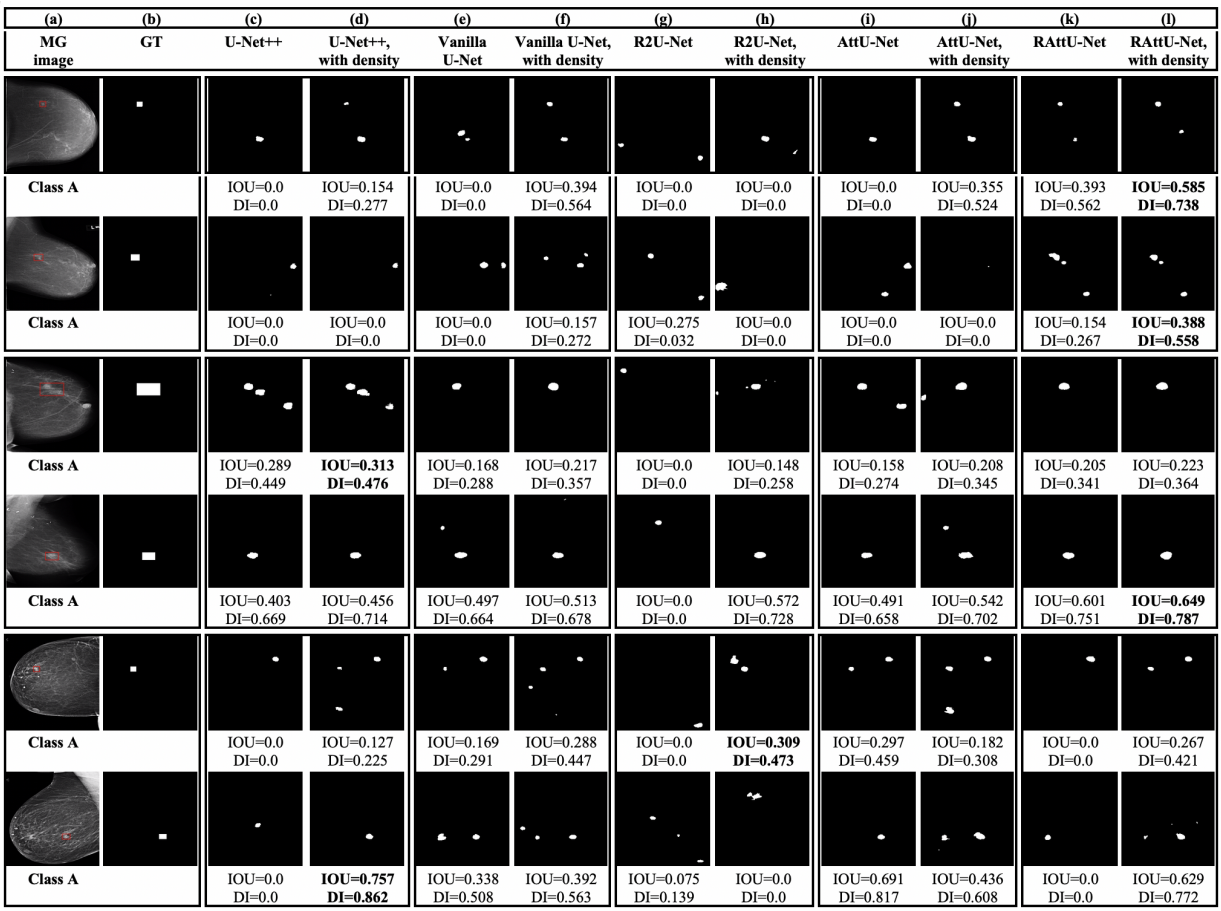


Fig. 10.8: (a) shows class A MG images of the UCHCDM data-set and its corresponding GT mask in (b); columns (c), (e), (g), (i), and (k) show the output of the trained models without pre-training on dense areas; (d), (f), (j), and (l) show the output of the trained models pre-trained on dense areas.

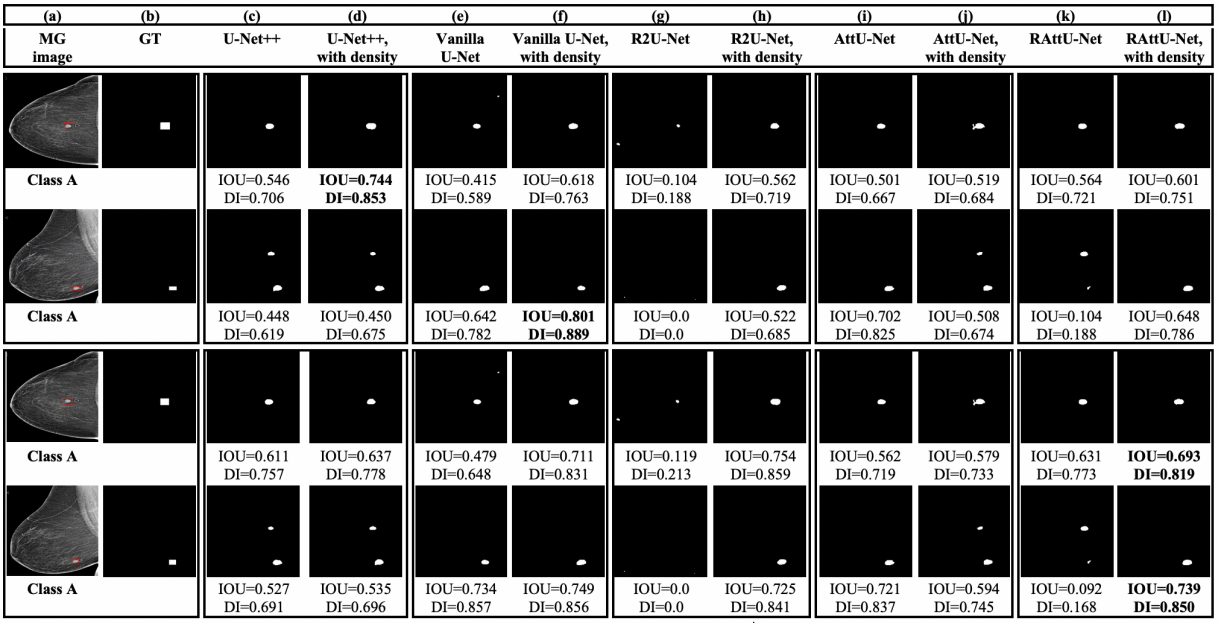


Fig. 10.9: (a) shows class A MG images of the UCHCDM data-set and its corresponding GT mask in (b); columns (c), (e), (g), (i), and (k) show the output of the trained models without pre-training on dense areas; (d), (f), (j), and (l) show the output of the trained models pre-trained on dense areas.

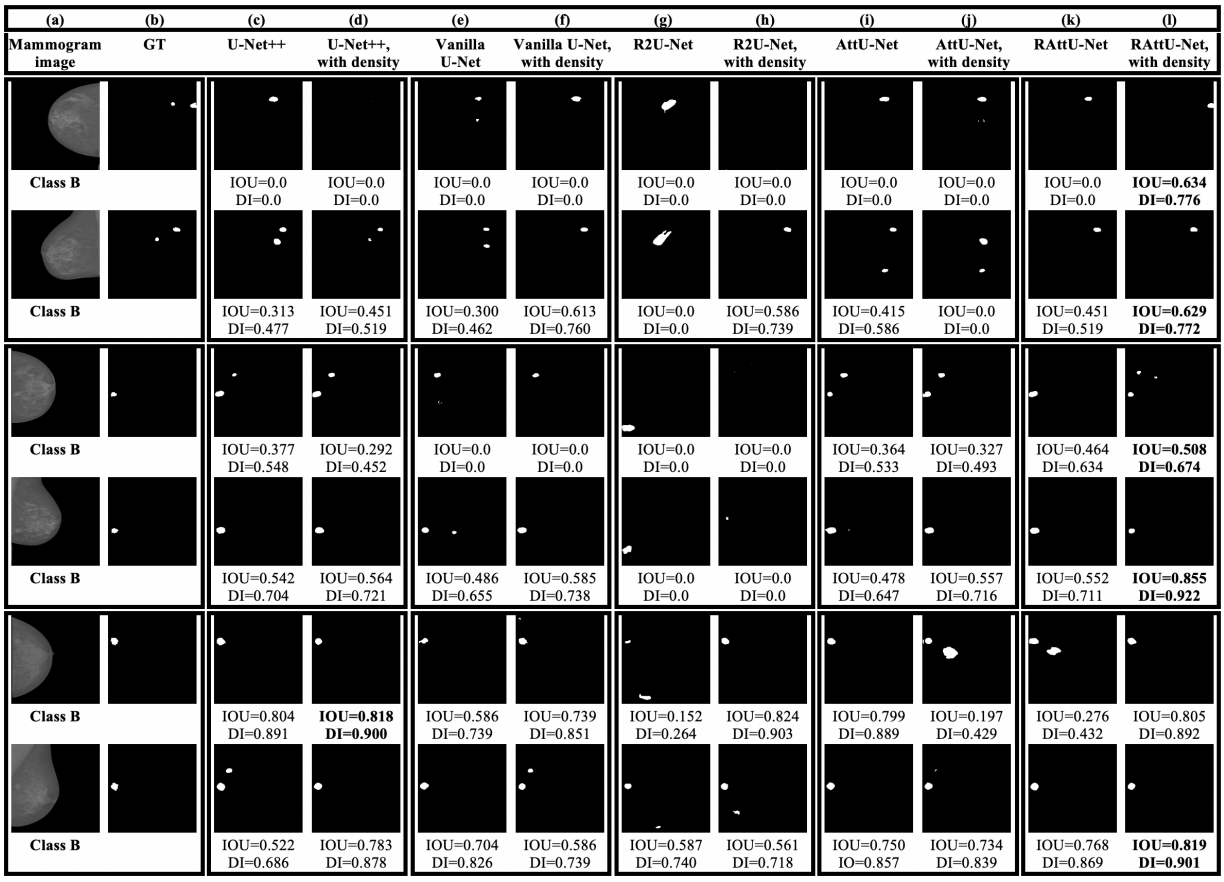


Fig. 10.10: (a) Shows class B MG images of the INbreast data-set and their corresponding GT mask in (b); columns (c), (e), (g), (i), and (k) show the output of the trained models without pre-training on dense areas; (d), (f), (j), and (l) show the output of the trained models pre-trained on dense areas.

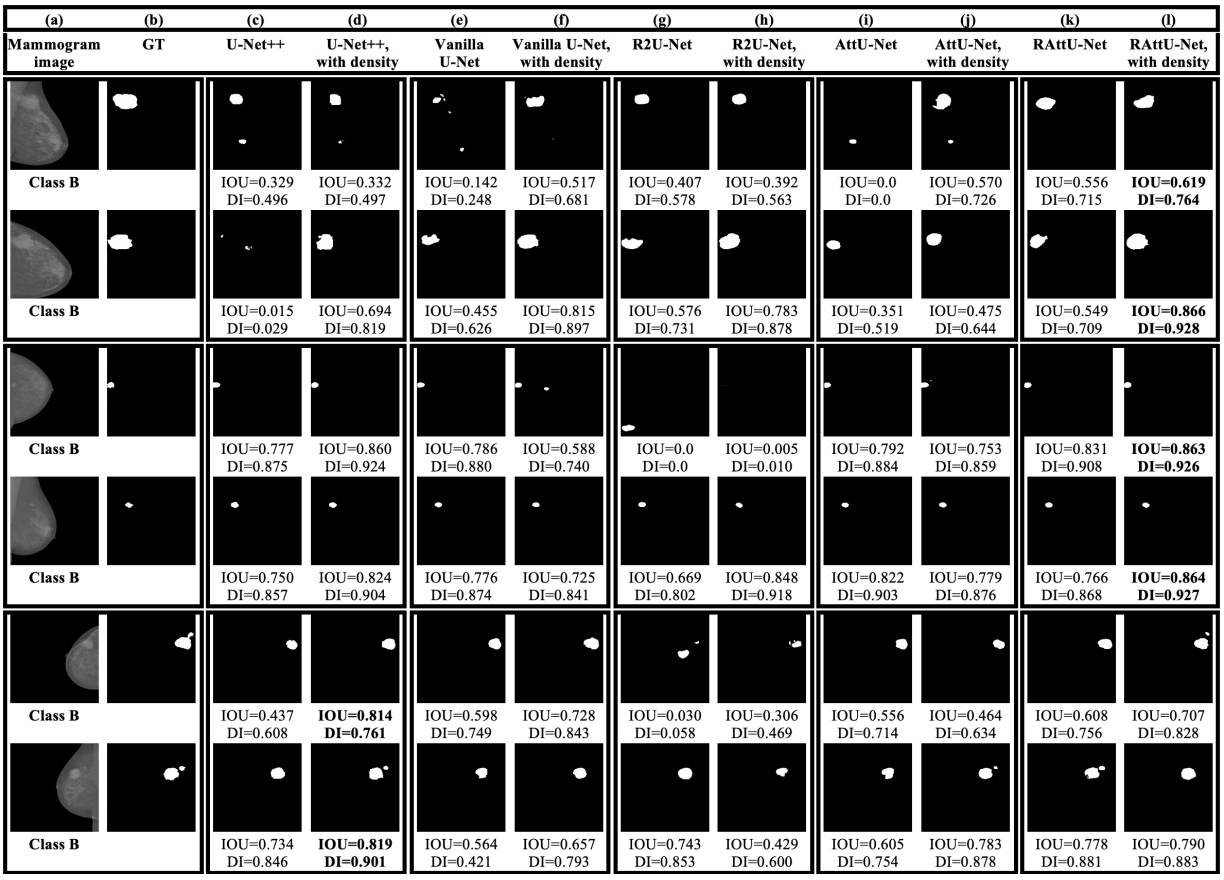


Fig. 10.11: (a) shows class B MG images of the INbreast data-set and its corresponding GT mask in (b); columns (c), (e), (g), (i), and (k) show the output of the trained models without pre-training on dense areas; (d), (f), (j), and (l) show the output of the trained models pre-trained on dense areas.

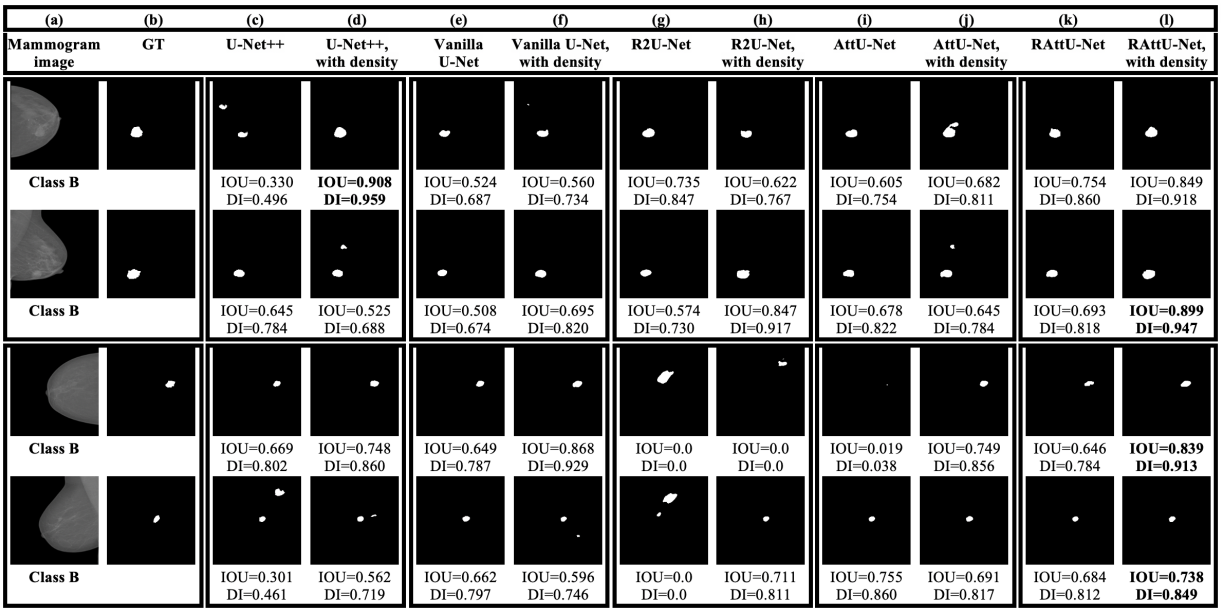


Fig. 10.12: (a) shows class B MG images of the INbreast data-set and its corresponding GT mask in (b); columns (c), (e), (g), (i), and (k) show the output of the trained models without pre-training on dense areas; (d), (f), (j), and (l) show the output of the trained models pre-trained on dense areas.

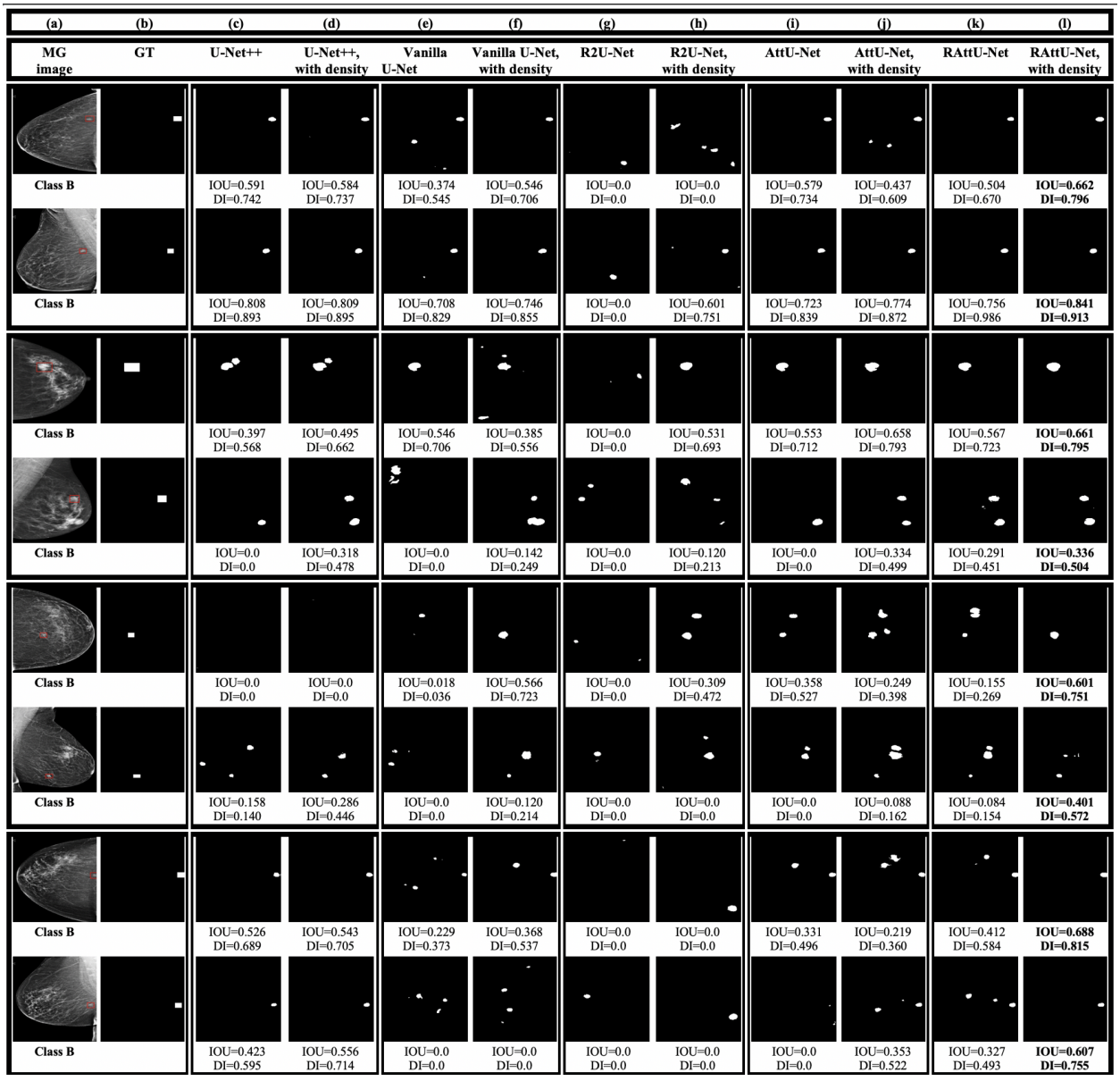


Fig. 10.13: (a) shows class B MG images of the UCHCDM data-set and its corresponding GT mask in (b); columns (c), (e), (g), (i), and (k) show the output of the trained models without pre-training on dense areas; (d), (f), (j), and (l) show the output of the trained models pre-trained on dense areas.

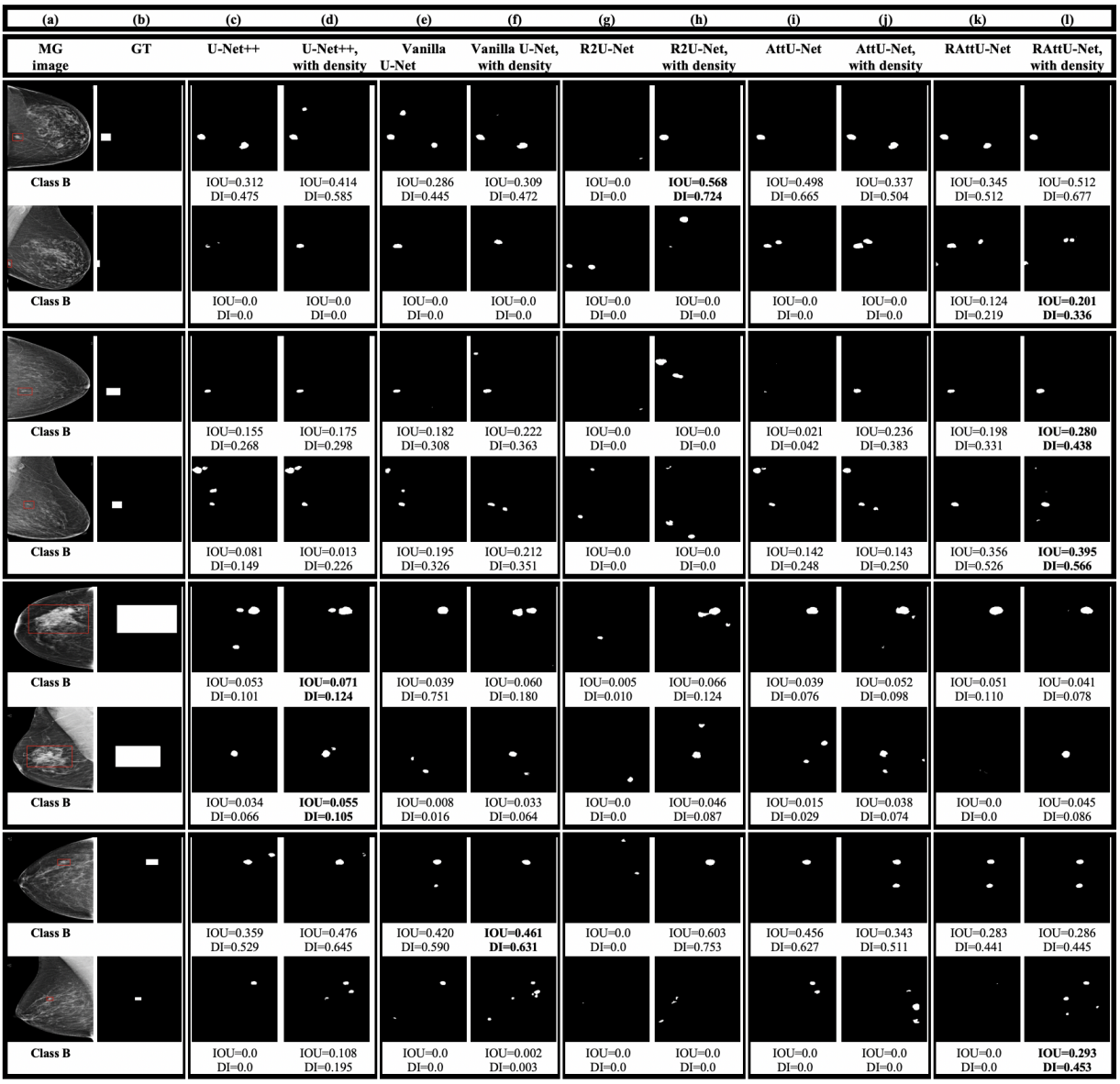


Fig. 10.14: (a) shows class B MG images of the UCHCDM data-set and its corresponding GT mask in (b); columns (c), (e), (g), (i), and (k) show the output of the trained models without pre-training on dense areas; (d), (f), (j), and (l) show the output of the trained models pre-trained on dense areas.

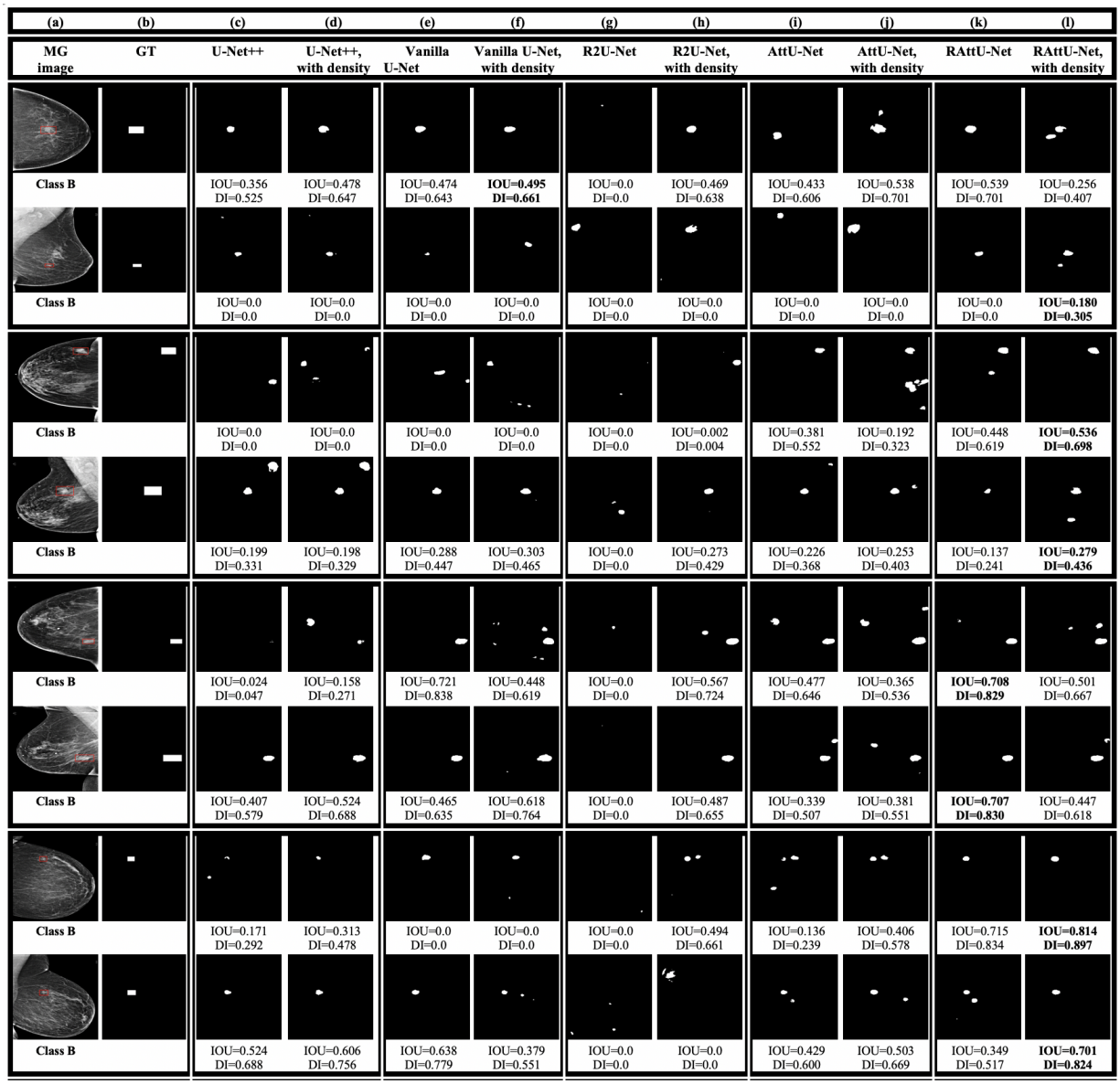


Fig. 10.15: (a) shows class B MG images of the UCHCDM data-set and its corresponding GT mask in (b); columns (c), (e), (g), (i), and (k) show the output of the trained models without pre-training on dense areas; (d), (f), (j), and (l) show the output of the trained models pre-trained on dense areas.

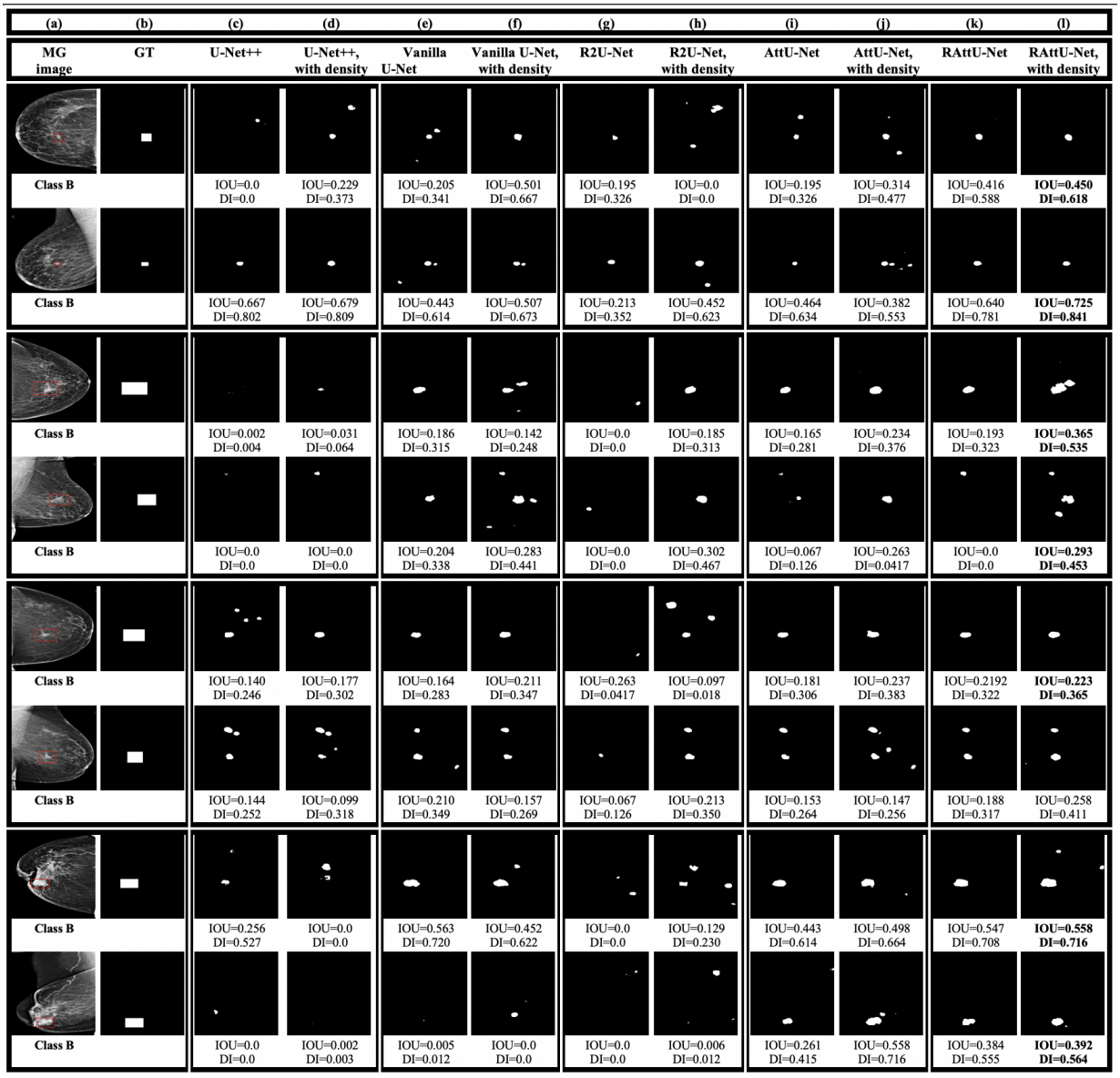


Fig. 10.16: (a) shows class B MG images of the UCHCDM data-set and its corresponding GT mask in (b); columns (c), (e), (g), (i), and (k) show the output of the trained models without pre-training on dense areas; (d), (f), (j), and (l) show the output of the trained models pre-trained on dense areas.

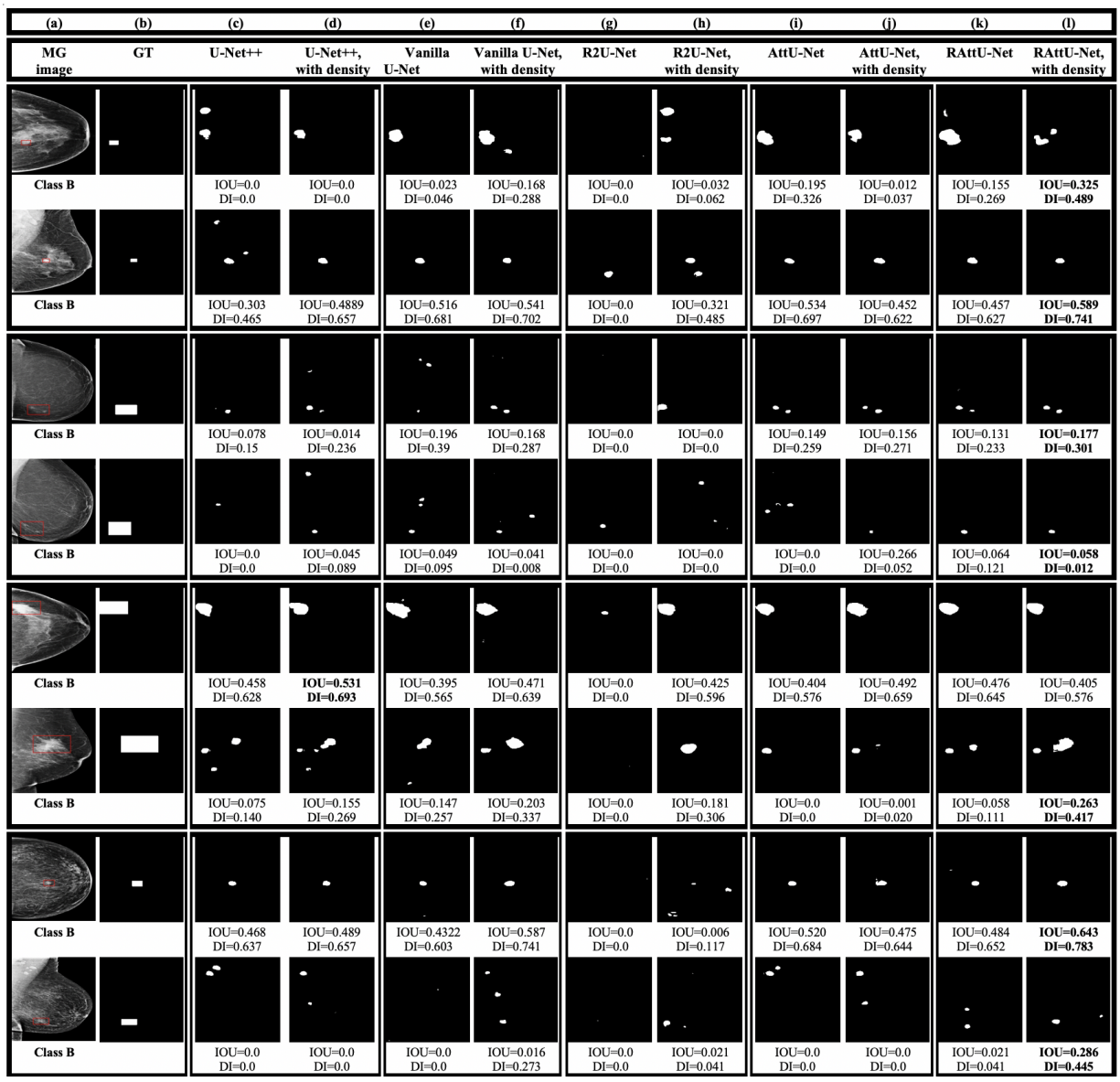


Fig. 10.17: (a) shows class B MG images of the UCHCDM data-set and its corresponding GT mask in (b); columns (c), (e), (g), (i), and (k) show the output of the trained models without pre-training on dense areas; (d), (f), (j), and (l) show the output of the trained models pre-trained on dense areas.

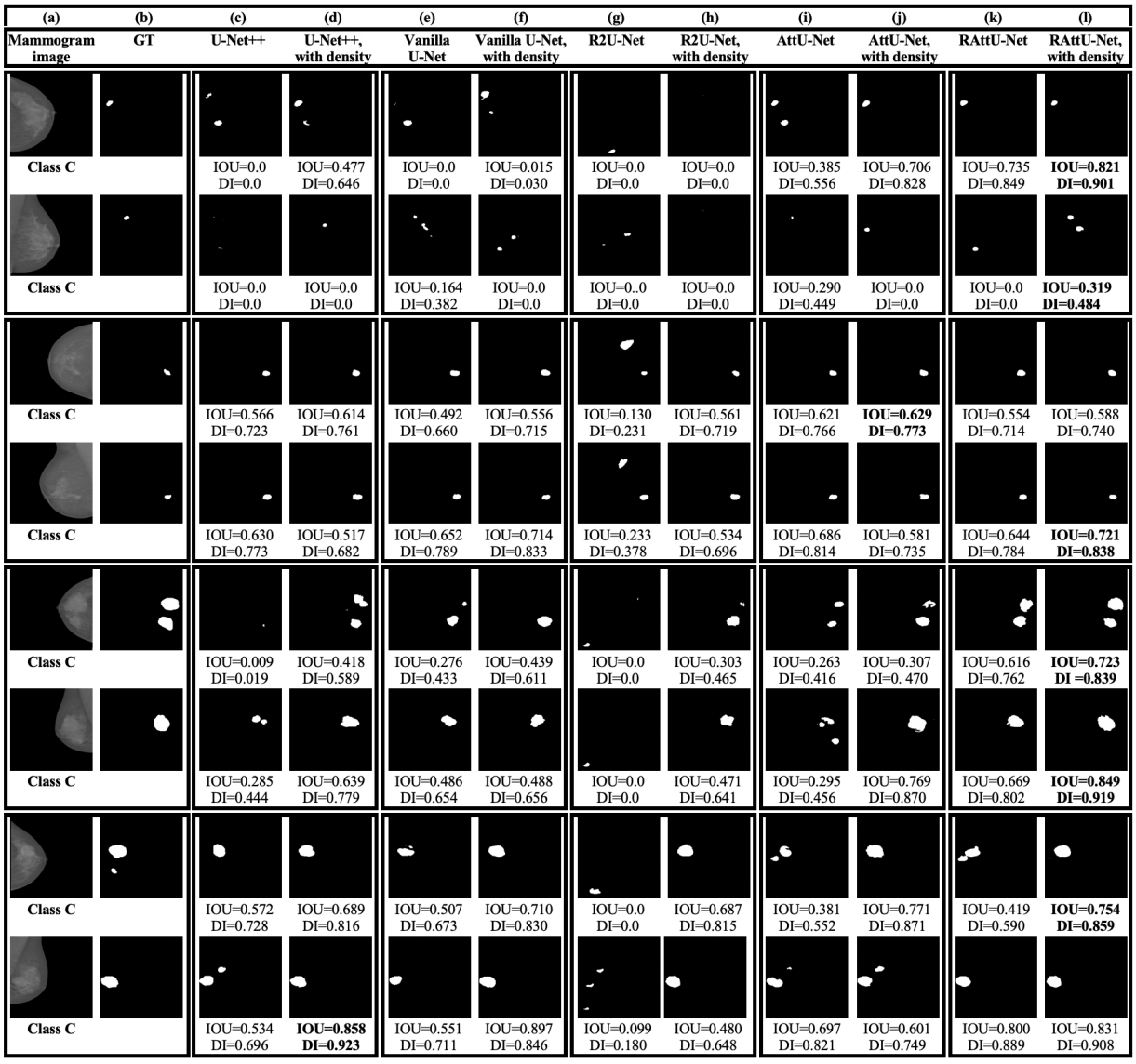


Fig. 10.18: (a) Shows class C MG images of the INbreast data-set and their corresponding GT mask in (b); columns (c), (e), (g), (i), and (k) show the output of the trained models without pre-training on dense areas; (d), (f), (j), and (l) show the output of the trained models pre-trained on dense areas.

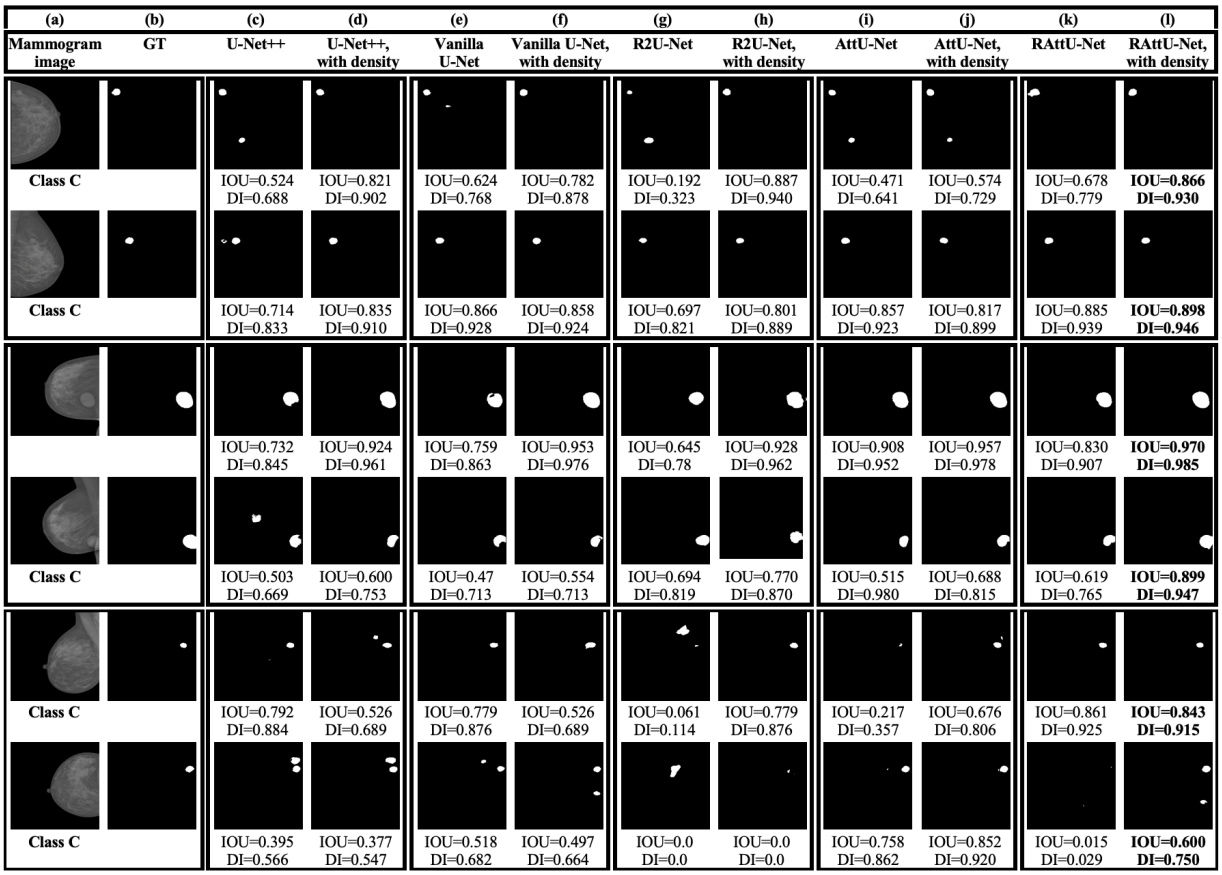


Fig. 10.19: (a) shows class C MG images of the INbreast data-set and its corresponding GT mask in (b); columns (c), (e), (g), (i), and (k) show the output of the trained models without pre-training on dense areas; (d), (f), (j), and (l) show the output of the trained models pre-trained on dense areas.

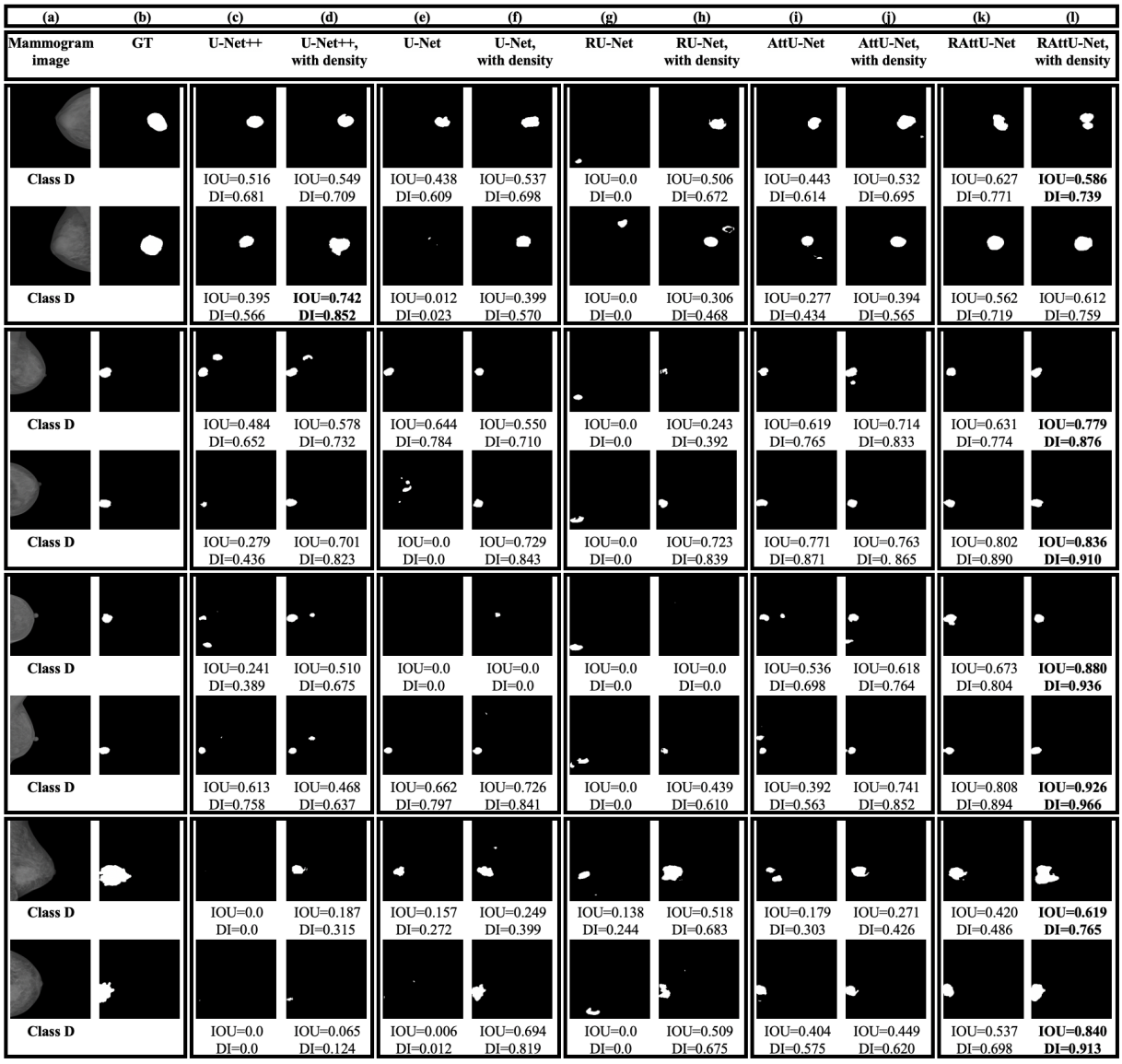


Fig. 10.20: (a) Shows class D MG images of the INbreast data-set and their corresponding GT mask in (b); columns (c), (e), (g), (i), and (k) show the output of the trained models without pre-training on dense areas; (d), (f), (j), and (l) show the output of the trained models pre-trained on dense areas.

10.5 Discussion

As we mentioned before, the MG images of class C (heterogeneously dense) and class D (extremely dense) are very challenging where the dense areas of the breast make it harder to find masses and obscure small masses. The proposed training strategy succeed to overcome these challenges (Figs. 10.18, **??**, and 10.20). This is because the generated density attention maps help the main CNN used for detecting lesions to pay attention more to these regions and predict a precise pixel-wise segmentation map of an input full MG image.

10.6 Conclusions

We compared the performance of the nested U-Net (U-Net++) model, vanilla U-Net, recurrent residual U-Net (R2U-Net), attention U-Net (AttU-Net), residual attention U-Net (RAttU-Net) that are trained from scratch on the segmentation of lesions with the corresponding models that are trained using the proposed training strategy. Furthermore, we trained the state-of-the-art, the Faster R-CNN model, and the Yolo model with our training data-sets from scratch to detect masses. We use the mean Dice Index (DI), mean intersection over union (IOU), and the inference time in seconds per image to measure the performance of each model. We show that using our training technique, the performance of the segmentation models increased in terms of DI and IOU. The RAttU-Net (pre-trained using the proposed training strategy) is superior in detecting masses in heterogeneous and dense MGs with a high mean IOU of 0.912 and 0.858, respectively than other models. Following it is the AttU-Net with a high IOU of 0.858 and 0.833, respectively. The RAttU-Net and U-Net++ show near mean IOU. The best model show improvement in detection of tumors with IOU increase from 0.923 to 0.950 (fatty class), from 0.887 to 0.928 (scattered class),

from 0.805 to 0.912 (heterogeneous class), and 0.757 to 0.858 (dense class).

Chapter 11

A Closer Look at Prior and Current MG Images using One-shot Image Recognition

11.1 Abstract

Radiologists compare MLO and CC views, look for asymmetry and evaluate changes with respect to prior year MGs. Findings that have not changed from older MGs are likely to be benign tumors and are signs of FP detection. In contrast, abnormal cells grow at a more rapid rate and begin to invade surrounding tissues, they form cancerous tumors. These cancerous tumors have different shapes and sizes than in prior images of the same patients. However, the appearance of normal breast tissues does not change much across the years.

We propose a deep learning One-shot Siamese network that is trained on both images and mimic radiologists in comparing prior and current MG images to improve the detection accuracy. The proposed Siamese network uses two identical sub-networks that have the same architecture and share the same parameters and weights.

The Siamese CNN approach achieved an AUC of 0.992 on the test UCHCDM data-set. We also compared the performance of the Siamese CNN model with a baseline CNN model. The accuracy of the classification of MG images increased by 14.22% when prior images were used to train the

DL CNN model. Using a baseline CNN model, the accuracy dropped from 0.932 to 0.816. Results show that using prior-year images in training DL models is a promising approach.

11.2 Background

Breast cancer is the second most common cause of cancer deaths among women in the USA after lung cancer. As of January 2019, there are more than 3.1 million women with a history of breast cancer in the USA [54]. Screening mammography is the main imaging modality used for early detection of breast abnormalities [34, 54]. The most important risk factors for the development of breast cancer are gender, age, and genetics. To reduce the risk of developing cancer among women, current guidelines from the American College of Radiology (ACR) and the National Comprehensive Cancer Network (NCCN) recommend screening mammography every year for women, beginning at age 40, as well as regular breast exams by radiologists to check for lumps. These follow-up screenings are done to see if there has been any change in an abnormality over time.

Studies have shown that mammogram (MG) images are the only method of breast imaging that consistently contributed to the decrease of breast cancer-related mortality [34, 54]. According to the American Cancer Society, the female breast cancer death rate declined by 40% from its maximum in 1989 to 2016 [67, 199] through early detection using mammography screening. Moreover, in the USA, full-field digital mammography (FFDM) has replaced conventional Screen-film mammography (SFM) [68, 205] to have better assessments.

Breast cancer can manifest on mammography in several ways: abnormal areas of lumps or tumors called masses, small white spots called calcifications (MCs), architectural distortion (AD), asymmetry, or any combination of these. Besides the primary aim of detecting breast cancers in

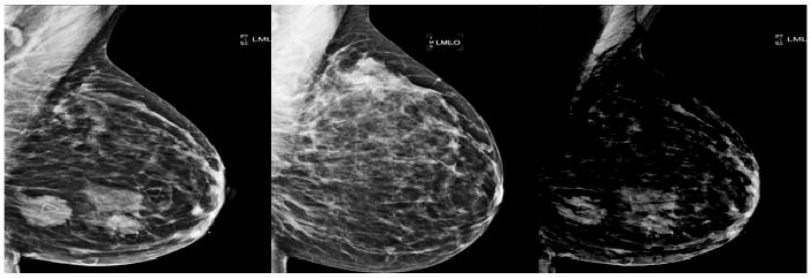
screening MGs, it is necessary to lower false-positive rates (FPRs) to avoid additional radiation exposure, and unnecessary biopsies [37].

Radiologists use several techniques to lower the FPRs, these techniques include double reading, prior-current year reading, multi-view reading, and modalities reading. The idea of double reading is to have two or more radiologists read the same MG [171], which leads to unnecessary increase costs and a similar cancer detection rate [172]. Beside double readings, radiologists compare current year MG images to prior MGs images acquired from the same breast for patients at different times. Prior-current year analysis (also called temporal analysis) of MG images are required for various reasons, such as monitoring the tumor growth or post-operative monitoring of healing. Tumors that have not changed from older MGs are likely to be benign tumors and are signs of FP detection. In contrast, abnormal cancerous cells grow at a more rapid rate and begin to invade surrounding tissues, as they form cancerous tumors. These cancerous tumors have different shapes and sizes than in prior images of the same patients [147]. Moreover, the presence of a suspicious mass, calcifications, or both, and a change in from prior MG, may suggest a high probability of breast cancer. Radiologists also use multi-view reading to reduce FPs by comparing the findings in both the craniocaudal (CC) and mediolateral oblique (MLO) views of the same breast [35, 79]. Yes, multi-view reading is important and helpful for radiology to see the findings from another view, however findings seen in one view might be unseen in the other view, this case is called architectural distortion (AD). In the multimodal analysis, images for the same breast are acquired using different modalities, e.g. X-ray and MRI. However, MRI is used for women at high risk of breast cancer, evaluate the extent of cancer following diagnosis, or further evaluate abnormalities seen on mammography.

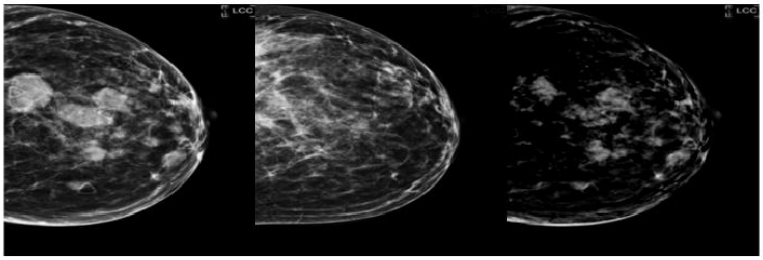
Studies have shown that the use of prior MGs in screening effectively reduces the number of FPs referrals [77]. In dense breasts, prior images allow radiologists to distinguish growing lesions from normal dense structures in the breast that somehow look suspicious. This help shows if any findings are new, or if they were already there on prior MGs. [231] shows that incorporating the temporal change information in mammography mass classification can improve the performance detection for CAD models. Automatic comparison between temporal MGs is still a difficult task for the current traditional CAD models because of the complexity of temporal MG registration [231]. These CAD models involve pre-processing, temporal image registration and temporal image subtraction to generate a difference image, and finally the analysis of the different images to obtain the changed and unchanged regions of the lesion. If miss-registration occurred, this causes subtraction artifacts which produce a false-positive result (see Fig. 11.1).

Recently, DL methods and in particular convolutional neural networks (CNNs) have achieved the-state-of-the-art performance for image classification and lesion detection for mammography [39, 57, 124, 127] and for medical applications in general [4, 10, 139]. DL can reduce the workload of radiologists and the cost associated with double reading. However, the performance of current DL systems still need improvements to enhance its accuracy [4, 10, 102, 138, 243]. Compared to the traditional CAD methods, DL methods offer automated learning of features estimated based on specific detection/segmentation/classification functions.

Sun et al. [212] proposed a method to use deep CNN for near-term breast cancer risk analysis by detecting a mass in two temporal MG screenings. In his study, every breast image is divided into 100 ROIs with 52×52 pixels, and each ROI were trained with the proposed CNN individually, and the final predictions of each case were based on the overall risk scores of all the 100 ROIs.



(a) MLO view.



(b) CC view.

Fig. 11.1: (a) and (b) show the conventional radiology method applied to the left images of same patient where the temporal (current and prior) MG images are subtracted to enhance areas where differences occur. However, without proper registration of temporal images pairs, the review of temporal image pairs may be seriously hampered by differences in acquisition and position changes.

Regardless of having a temporal data-set, Sun et al. did not use prior images to reduce FPR in current images.

Carneiro et al. [40] addressed the classification of mass(es) using a multi-view pre-trained CNN.

Their model classifies a full MG by extracting features from each view of the breast (train a separate CNN for each view) and combining these features in a joint CNN model to output a prediction that estimates the patient's risk of developing breast cancer. Carneiro et al [41] following his work [40] build a fully automated pre-trained CNN for detecting masses and MCs in multi-view images reaching AUC of 0.9 for the 2-class benign versus the malignant problem. Dhungel et al. [58] trained a multi-view deep residual neural network (mResNet) for the fully automated classification of MGs as either malignant or normal/benign. Dhungel et al. [58] show that the combination of both views with the automatically generated lesion segmentation masks produces a reasonably accurate classification into malignant or normal/benign, with an AUC of 0.8.

Deep convolutional neural networks (DCNNs) are the state-of-the-art methods used for image classification and retrieval tasks. The main challenge for researchers is the need of collecting large annotated training data-sets. In applications using medical images, collecting this much data is sometimes not feasible. We propose a Siamese network with one-shot learning, that aims to solve this problem. To train such a DL network to compare temporal MG images (prior-current year images), we first require temporal MG images of each patient. The proposed network is not learning to classify a temporal MG image directly to any of the output classes (normal, cancer). Rather, the proposed network instead takes an extra reference image (prior image of the same view) of the same patient as input and will produce a similarity score that expresses how similar they are. The dissimilarity between two pairs of temporal MG images shows a risk factor of developing cancer within the tissues. The proposed Siamese network is a type of DL network that uses two identical sub-networks that have the same architecture and share the same parameters and weights. Some common applications for Siamese networks include facial recognition, signature

verification. Siamese networks are particularly useful in cases where there are large numbers of classes with small numbers of observations of each, which is our case here. In such cases, there is not enough data to train a DL to classify images into normal, cancer classes. If new temporal images are introduced to the Siamese network, the network determines if two images are similar or dissimilar. Siamese networks perform well in these tasks because their shared weights mean there are fewer parameters to learn during training and they can produce good results with a relatively small amount of training data.

The performance of the Siamese model is compared to a baseline architecture without integrating prior-year images using Area Under the ROC Curve (AUC). The Siamese network is trained on temporal images from the public breast cancer digital repository (BCDR) and the private full-field digital repository UCHCDM from the University of Connecticut health care center (UCHC).

11.3 Material and Methods

11.3.1 Siamese Network using One-Shot Recognition

We propose a deep learning Siamese network using two identical CNNs sharing the same weights (see Fig. 11.2). In our experiment, each identical CNN is composed of three blocks of convolution layers (conv.), ReLU, and Max-pooling followed by a Conv. connected to a fully-connected (FC) layer with a sigmoid function. This FC layer produces the feature vectors that will be fused by the weighed distance layer. The output is fed to a final FC layer that outputs a value between 1 and 0 (dissimilar or similar images).

To assess the best architecture, we used a mini-batch size of 4 and the Bayesian hyperparameter optimization to find the best parameters used in our architecture.

- Layer-wise learning rates (search from 10^{-5}: 10^{-1}).
- Layer-wise momentum (μ) (search from 0: 1).
- Layer-wise L2-regularization weight (λ) (from 0: 10^{-1}).
- Filter size from 3×3 to 10×10.
- Filter numbers from 16 to 256 (using multipliers of 16).
- Number of units in the fully connected layer from 128 to 4096.

We used a 25% dropout for regularization to reduce overfitting. The rectified linear units (ReLU) was applied as the activation function after all the Conv. layers except for the last Conv. layer, and we used L2-Regularization in each layer and the Stochastic Gradient Descent with momentum as an optimizer. The units in the final Conv. layer are flattened into a single vector. This Conv. layer is followed by an FC layer, and then one more layer computing the distance metric between each CNN, which is given to a single sigmoidal output unit (see Figs. 11.2 and 11.3).

In the proposed supervised learning approach, our data-set contains pairs of (X_i, ($D_i(x_1, x_2)$) where X_i is input pair of images and ($D_i(x_1, x_2)$) is the output similarity score. $D_i(x_1, x_2) = 0$; if both images are similar, $D_i(x_1, x_2) = 1$; if both images are dissimilar.

The network employs a pair of images x_1 and x_2 as input and is passed through the Siamese network to generate a fixed-length feature vector for each ($h(x_1)$, $h(x_2)$). The function ($h(x_1)$, $h(x_2)$) represents the embedding features from each pair of images extracted by the Siamese network. If the two input MG images (prior and current) has no change in texture, then their feature vectors must also be similar, while if the two input images belong to two different views or have changed in the texture (i.e. masses or calcification developed), then their feature vectors will also be different.

The similarity between pair of images is calculated by the Euclidean distance $D(x_1, x_2)$, which is element-wise absolute difference between the two feature vectors $h(x_1)$, $h(x_2)$ (Equation 11.1). We establish the constructive loss function (L) [88] (Equation 11.2), which minimizes the Euclidean distance between the features of similar pairs and maximize it for dissimilar pairs.

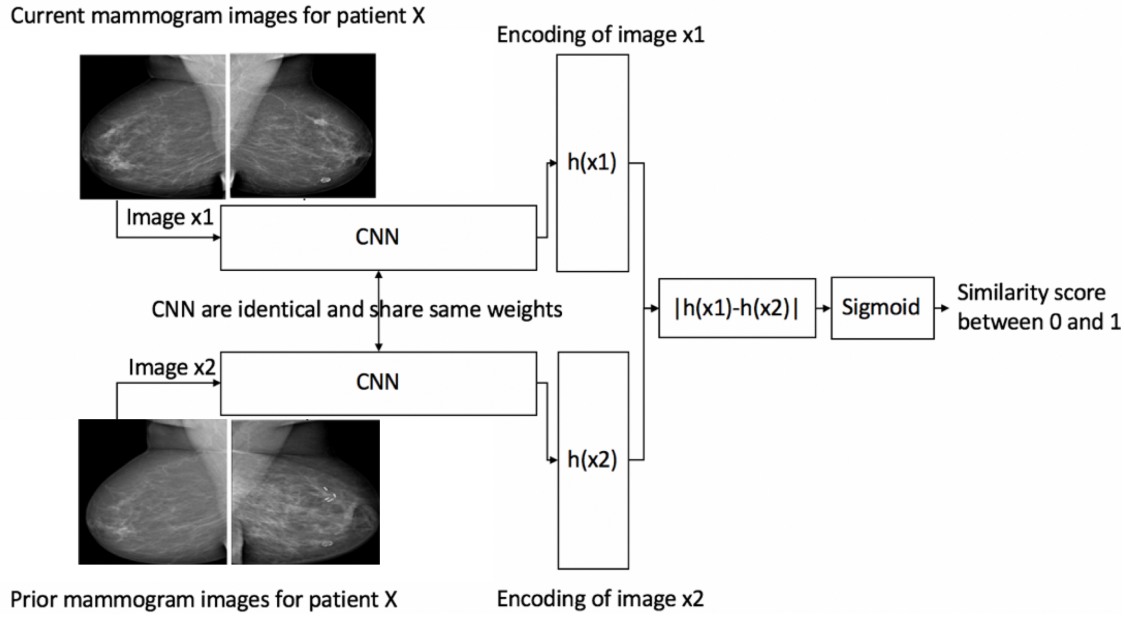


Fig. 11.2: General architecture of the proposed Siamese network. The network is first pre-trained on patch pairs cancer and non-cancer images. Matching pair of images and non-matching images. Then the network is fine-tuned with pairs of full MG images from current year and prior year views. The feature network, consists of interchanging layers of convolutions and pooling, share parameters between paths.

$$D(x_1, x_2) = \|h(x_1) - h(x_2)\|^2, \tag{11.1}$$

$$L(x_1, x_2; m) = \frac{1}{2} \times Y \times D(x_1, x_2) + \frac{1}{2} \times (1 - Y) \times max(0, m - D(x_1, x_2)), \tag{11.2}$$

Where Y is a binary label of the input images x_1 and x_1. For x_i pair of images,$Y = 1$ indicates that the image pairs are dissimilar and $Y = 0$ shows that the image pairs are similar. The parameter

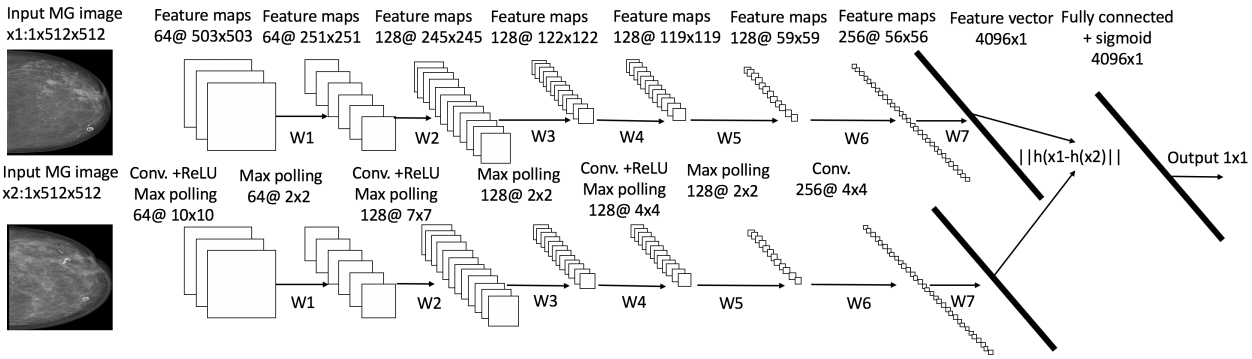


Fig. 11.3: Detailed architecture of the proposed Siamese network.

m is defined as the margin threshold between similar and dissimilar pairs. Equation 11.2 calculates loss per sample during training. When similar image pair, which has $Y = 0$, is fed to the network, the right-hand side of the addition section of Equation 11.2 will be equal to zero, thus $\frac{1}{2} \times Y \times D(x_1, x_2) = 0$ and the loss becomes equal to the part containing the positive pair distance between the embedding of two similar images $(\frac{1}{2} \times (1 - Y) \times max(0, m - D(x_1, x_2))$. If the image pair has $Y = 0$ but the network output a feature pair of vectors whose $D(x_1, x_2)$ is completely dissimilar thus $D(x_1, x_2)$ has value greater than m, then the value of the loss function is maximized to zero else if the images have similar $D(x_1, x_2)$, then the gradient descent reduces the distance between them which is learned by the network.

On the other hand, when two dissimilar images, $Y = 1$, are fed to the network, the left-hand side of the addition turns to zero, thus $\frac{1}{2} \times (1 - Y) \times max(0, m - D(x_1, x_2)) = 0$ and the remaining part of the equation $(\frac{1}{2} \times Y \times D(x_1, x_2))$ works as the loss function. Thus the variable m is the margin of separation between dissimilar and similar samples and is decided empirically. When m is large, it pushes dissimilar and similar images further apart thus acting as a margin, in our experiments $m = 1$.

We follow the same initialization strategy introduced by Koch et al. [120]. The learning rate (Lr)

is defined layer-wise and it is decayed uniformly by 0.1% each epoch (T), thus $Lr(epoch = T) = 0.99 \times Lr(T - 1)$). We trained each network for a maximum of $T_{max} = 300$ epochs. In every layer, the momentum μ is fixed at 0.5 and it is increased linearly each epoch until reaching a value mu. For each image pair x_1, x_2, data augmentation was used with affine distortions T = ((rotations:θ), (translations: t_x, t_y) and (zoom: ρ_x, ρ_y). With $\theta \in [-10.0, 10.0]$, ρ_x, $\rho_y \in [-0.3, 0.3]$, and $t_x, t_y \in [-1, 1]$).

In our data-set, we have the following similar MG images from same view with [(N_C, N_P), (B_C, B_P), (M_C, M_p] and dissimilar images from the same view with [M_C, N_P), (B_C, N_P)], where B refers to benign, M refers to malignant, N refers to normal, C refers to current, and P refers to prior. The UCONN data-set has mainly these cases only [(N_C and (M_C, N_P)]. However, the BCDR data-set has all the cases [(N_C, N_P), (B_C, B_P), (M_C, M_p), (M_C, N_P), (B_C, N_P)].

Notice that this network is not learning to classify an image directly to any output classes. Rather, it is learning a similarity function, which takes two images as input and expresses how similar they are. Our target that the Siamese network recognizes the development of tumor in the breast and can show dissimilar output for these cases (M_C, N_P), (B_C, N_P).

Training Strategy for Siamese Network

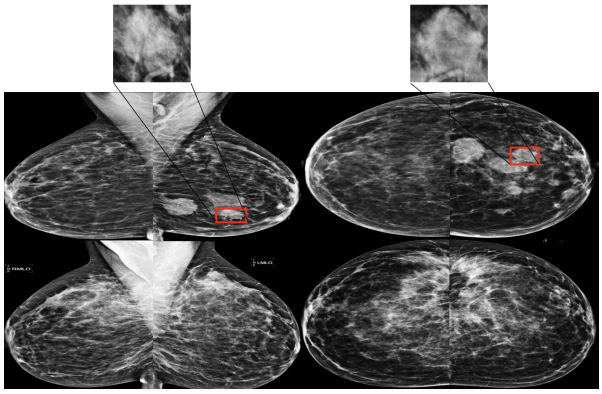
Before we train our Siamese network for our recognition task of temporal similar or dissimilar images, we first pre-train it for classification task with all patches that has tumors centered in the image (as cancer class) and all patches of images that are normal (as normal class). By this way the network learn to differentiate between the textures that are normal and textures that has tumor. We did that pre-training for two reasons, the first one is because the prior normal image and the

current malignant or benign image have large similarity except in small areas which have tumor. These areas can vary in size and can be very small to the point that the network can consider the two images as similar but it is not the case here. Current and prior images are considered different because of small area(s) that developed tumor. Small tumors will make the network struggle giving similar output while the labels are dissimilar. The proposed Siamese network is pre-trained on augmented patches (normal and tumors of different sizes, shapes and margins), then the weights of this network are used as initialization for the Siamese network for the whole image recognition task. So instead of using a CNN architecture for the recognition task with the ImageNet pre-trained weight as in these architectures [51, 120], we used the weights of the network trained on patches of normal and cancer images.

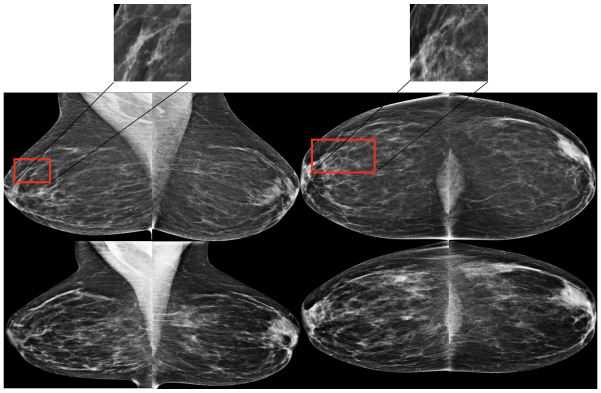
Data-sets for Siamese Network

This study is conducted using 211 cases (184 cancerous versus 27 healthy) of originally digital MGs collected at University of Connecticut Health Center (UCHC), called UCHC DigiMammo (UCHCDM) database [9, 253]. Each case includes four MGs (two views [CC and MLO] from two sides) imaged at two different times, referred as current (C) and prior (P) exam or scan (see Fig. 11.4). Each mammographic image has a large size (e.g., 3328×4096 pixels) and 12 bits per pixel. UCHCDM images are originally stored in DICOM format. All mammographic images are ground truth annotated files with a description of known pathology (normal (healthy), mass, AD, calcification) and circled at the locations of cancers (if any) on a separate key ground truth image. In 184 cancerous cases, there are labels for clarifications, architectural distortion (ADs) and masses annotated by our expert radiologists. The mammographic images are first pre-processed by employing normalization such as scaling image intensity to the range of $[0, 255]$. From the

UCHCDM data-set, we extracted 368 current cancerous images that have 368 prior normal images, and we extracted 476 current normal images that have 476 prior normal images (see Table 11.1).



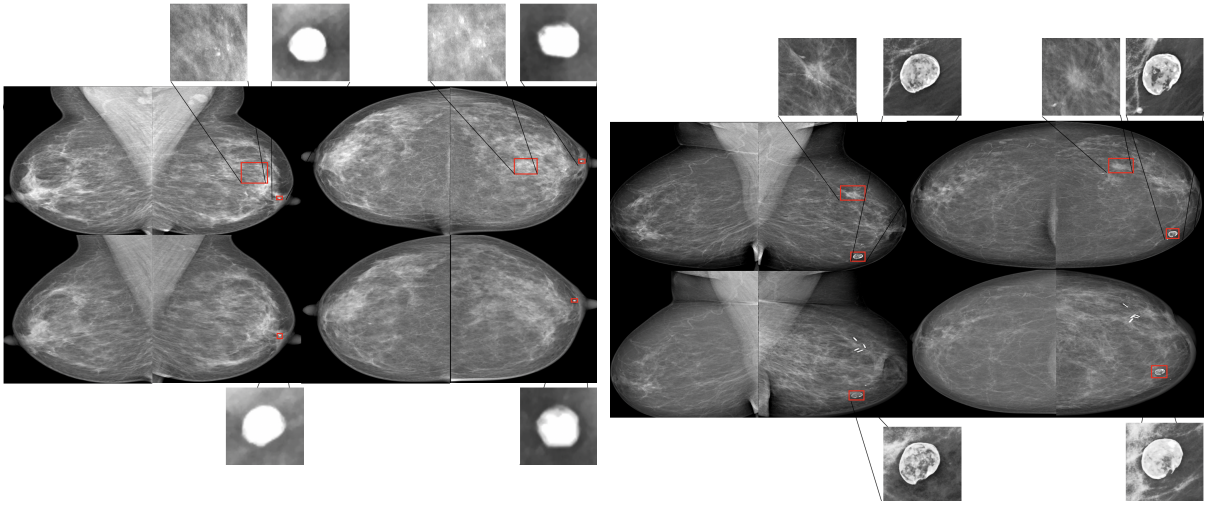
(a) Patient has heterogeneously dense breast. The left breast of C exam has malignant (BI-RADS 5) invasive ductal carcinoma mass, while the right breast of C exam has no lesions (normal). The right and left breasts of the P exam of same patient has BI-RADS 1 (normal). The C and P exams are scanned 6 years apart.



(b) Patient has breast density with scattered areas of fibroglandular. The right breast of C exam has malignant (BI-RADS 4) pleomorphic calcifications with segmental distribution, while the left breast of C exam has no lesions (normal). The right and left breasts of the P exam of same patient has BI-RADS 1. The C and P exams are scanned 6 years apart.

Fig. 11.4: Each case in (a) and (b) includes 4 mammograms (two views [CC and MLO] from two sides) imaged at two different times, the upper row in each case refers to current (C) exam and the lower row in each case refers to prior (P) exam. (a) and (b) are temporal cases from the UCHCDM data-set.

We also used some cases from the Breast cancer digital repository (BCDR) [142] that have temporal images for same patients (see Fig. 11.5). The BCDR database is subdivided in two different repositories: 1- a film mammography-based Repository (BCDR-FM) and 2- a full field digital mammography-based reposi-



(a) Patient has heterogeneously dense breast. The left breast of C and P exams has benign mass with round shape and circumscribed margins, while the right breast of C and P exams has no lesions (normal). Benign calcification are detected in the C exam of left breast. The C and P exams are scanned 1 year apart.

(b) Patient has scattered areas of fibroglandular. The left breast of C and P exams has benign mass with round shape and circumscribed margins, while the right breast of C and P exams has no lesions. Malignant mass of irregular shape and spiculated margins are detected in the C exam of left breast. The C and P exams are scanned 1 year apart.

Fig. 11.5: Each case in (c) and (d) includes 4 mammograms (two views [CC and MLO] from two sides) imaged at two different times, the upper row in each case refers to current (C) exam and the lower row in each case refers to prior (P) exam. (c) and (d) are temporal cases from the BCDR data-set.

tory (BCDR-DM). BCDR has patients' cases annotated by expert radiologists containing clinical data (detected anomalies, breast density, BI-RADS classification, etc.), lesions outlines, and image-based features computed from CC and MLO mammography image views. The MLO and CC images of the BCDR-FM data-set are grey-level digitized MGs with a resolution of 720×1168 pixels and a bit depth of 8 bits per

pixel, saved in the TIFF format. The MLO and CC images of the BCDR-DM are grey-level MGs with a resolution of 3328×4084 or 2560×3328 pixels, depending on the compression plate used in the acquisition (according to the breast size of the patient). The bit depth is 14 bits per pixel and the images are saved in the TIFF format.

We extracted from the BCDR data-set all the temporal images for each patient if available. We extracted from the BCDR data-set 257 current cancerous images that have 257 prior cancerous images. We did not use the current-prior normal breast side of the cancerous patients from the BCDR data-set.

All images related to same patient are grouped in the training data-set, validation data-set or the test data-set. We tested our model on 37 patients from the UCHCDM data-set that have current cancer and prior normal (for one side), and have current normal and prior normal (for the other side). In total we have 1300 training MG images, 388 validation MG images, and 296 pair of test images (see Table 11.1).

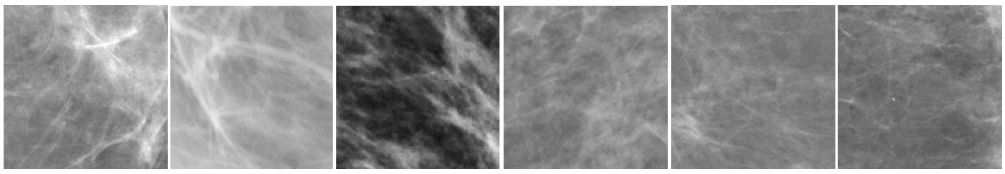
For pre-training our network, we used patches of normal and cancerous images from the DDSM, and the INbreast database (see Fig. 11.6). Cancer patches were extracted around the center-of-mass of the annotated lesion. Normal patches were randomly selected within the breast from the normal cases.

DDSM database comprises approximately 2,620 cases each containing both MLO and CC views of each breast. It contains digitized screen film mammography (SFM) images. The DDSM database contains the pixel-level annotation for the ROIs and their pathology confirmed labeling benign or malignant. It further contains the type of lesion, such as calcification or mass. Most MGs contain only one ROI while a small portion contain more than one ROIs. We converted the lossless jpeg images to png format, and re-scaled the pixel values to the 0-255 range. For the training of our model, we concerned only the images with masses and calcifications, having 913 of benign findings and 849 of malignant findings, In total, 1,762 patches.

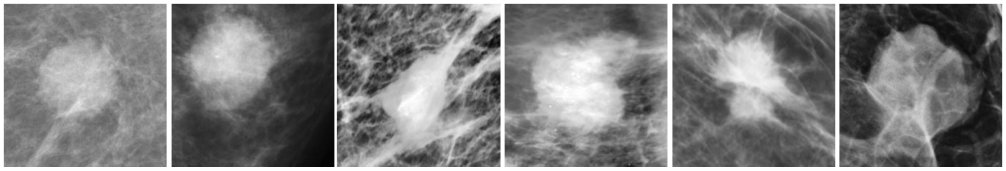
The INbreast comprises FFDM database. It has a total of 410 images, 116 mass annotated images and 343 calcification annotated images.

In total, we extracted balanced patches of normal and cancerous images, thus we used all 2,220 cancer

patches and 2,220 normal patches to pre-train our proposed model (see Table 11.1).



(a) Example of normal (healthy) patches used to pre-train our model.



(b) Example of cancerous patches used to pre-train our model.

Fig. 11.6: For pre-train our models, we use normal patches (a) and cancerous patches (b). Images shown are extracted from the (DDSM, BCDR, and INbreast) data-sets.

Table 11.1: The distribution of the data-sets used in training and testing our proposed models.

Data-set	**Purpose**	**#Images**	**Total**
DDSM (patches)	Pre-training	1,762	2,221
INbreast (patches)	Pre-training	459	
UCHCDM (full image)	Training	1,300	1,710
BCDR (full image)	Training	410	
UCHCDM (full image)	Validation	92	194
BCDR (full image)	Validation	102	
UCHCDM (full image)	Test	296	296

Pre-processing and Data-augmentation for Siamese Network

For our training, each original image is quantized to 8 bits per pixel (256 gray level). The dark background areas are cropped off, which leaves the breast area, after that, we apply the contrast limited adaptive histogram equalization (CLAHE) to enhance the contrast of the images. All MGs are re-scaled to the 0-255 range and converted into png format and re-sized into 512×512.

For the full pair of images, we artificially enlarged the data-set by using data augmentation. For each image pair x_1, x_2, data augmentation was used with affine distortions T = ((rotations:θ), (translations: t_x, t_y) and (zoom: ρ_x, ρ_y). With $\theta \in [-10.0, 10.0]$, ρ_x, $\rho_y \in [-0.3, 0.3]$, and t_x, $t_y \in [-1, 1]$). We used same the augmentation technique with the patches of images for pre-training our Siamese model.

11.3.2 Baseline DL architectures

The performance of the proposed Siamese network, which incorporate prior-current year in its training, is compared to a baseline architectures without integrating prior year images using Area Under the ROC Curve (AUC). In this work we are using the whole semantic segmentation nested U-NET model (U-Net++) and a proposed whole image classifier as our baseline DL architectures.

Whole Image Semantic Segmentation using U-Net++

The U-Net++ DL model with dense skip connections is adopted as our semantic segmentation model to learn multi-scale and different semantic levels of visual features representations in our work. U-Net++ consists of U-Nets of varying depths whose decoders are densely connected at the same resolution via the redesigned skip pathways. The U-Net++ architecture is an improved version of the U-Net as it embeds U-Nets of varying depths in its architecture, enabling improved segmentation performance for varying size lesions over the fixed-depth U-Net. The skip connections in U-Net++, enable flexible feature fusion in decoders in comparison to the restrictive skip connections in U-Net that require fusion of only same-scale feature maps.

For example, in U-Net++, node $X^{0,4}$ receives the skip connections from all previous convolution units at the same level, namely $X^{0,0}$, $X^{0,1}$, $X^{0,2}$ and $X^{0,3}$, while in U-Net only one skip connection is applied from node $X^{0,0}$. In such a way, the semantic levels of the encoder feature maps are closer to those in the corresponding decoder part. Nodes $X^{i,j}$ at level j of =0 receive only one input from a previous down-sampling layer, while nodes at level $j > 0$ receive $j + 1$ inputs from both the skip pathways and the up-sampling layer.

We used residual blocks instead of regular convolution block, which facilitates better convergence abilities for our deep networks. A 2-D conv.layer is implemented first, which is followed by a BN layer. Then, a further conv. and BN layer is applied. Finally, the output will be generated by adding the outputs from the second BN layer and the first Conv2D layer. We used the ReLU activation function after each conv. layer. A 1×1 convolution with C kernels followed by a Sigmoid activation function is appended to the outputs from nodes $X^{0,1}, X^{0,2}, X^{0,3}$, and $X^{0,4}$ to obtain results $Y^{0,1}, Y^{0,2}, Y^{0,3}, Y^{0,4}$. $X^{0,5} = Y^{0,1} \oplus Y^{0,2} \oplus Y^{0,3} \oplus Y^{0,4}$. $X^{0,5}$ is followed by a sigmoid layer, and the fusion output $Y^{0,5}$ is generated. Where C=2, the number of classes observed in the given data-set (background, lesion). A hybrid segmentation loss is used which consist of pixel-wise cross-entropy loss and soft dice-coefficient loss for each semantic scale. Mathematically, the hybrid loss is defined as:

$$\ell_i(Y, P) = \frac{1}{-N} \sum_{C=1}^{C} \sum_{n=1}^{N} y_{n,c} \times log p_{n,c} + \frac{2 y_{n,c} \times p_{n,c}}{y_{n,c}^2 + p_{n,c}^2}, \tag{11.3}$$

Where $y_{n,c} \in Y$ and $p_{n,c} \in P$ denote the target labels and predicted probabilities for class c and nth pixel in the batch, N indicates the number of pixels within one batch. The overall loss function for U-Net++ is then defined as the weighted summation of the hybrid loss from each individual decoders:

$$\ell = \sum_{i=i}^{5} \omega_i \times \ell_i(Y, P), \tag{11.4}$$

Where the corresponding weights are denoted as ω_i(i=1, 2, 3, 4, 5), and i is index of the decoder. To improve segmentation performance and weaken the effect of class imbalance problems, dice coefficient loss is usually applied in semantic segmentation tasks.

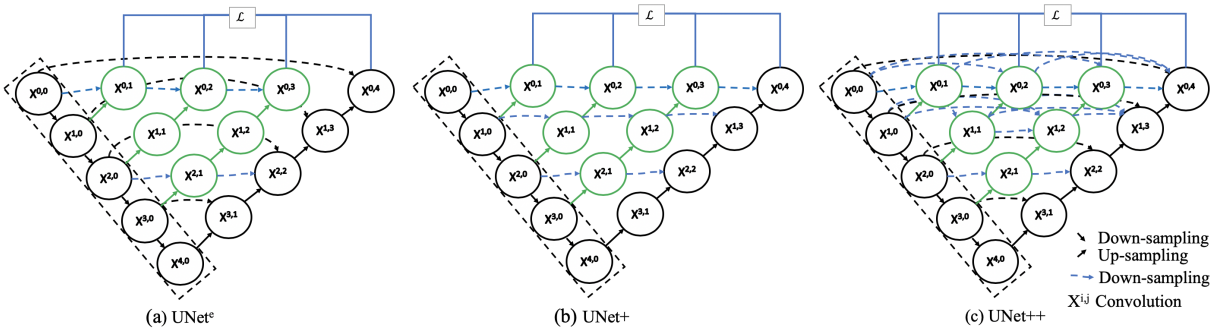


Fig. 11.7: Nested U-Net or U-Net++ consists of an encoder and decoder that are connected through a series of nested dense Conv. blocks. U-Net++ bridge the semantic gap between the feature maps of the encoder and decoder prior to fusion. For example, the semantic gap between $(X^{0,0}, X^{1,3})$ is bridged using a dense convolution block with three convolution layers. Black diagrams indicates the original U-Net, green and blue show dense Conv. blocks on the skip pathways. green, and blue components distinguish U-Net++ from U-Net.

Data-sets for Semantic Segmentation U-Net++ Model

We trained our CNN models on the publicly available data-sets, DDSM [93], BCDR [142], and the publicly available INbreast data-set [38] and tested it on the private UCHCDM data-set [9,253]. Each MG image has been annotated based on their density derived from the American College of Radiology's (ACR) Breast Imaging Reporting and Data System (BI-RADS) [29]. For each MG image, its density in ACR standard scale is given as one of these categories: class A: fatty, class B: scattered, class C: heterogeneous dense, and class D: dense [4, 29]. We grouped the BI-RADS multi-class assessment into benign and malignant classes. In this study, we categorized 1,133 MG images with BI-RADS $\in$ {2, 3} as benign, and 2,139 MG images with BI-RADS $\in$ {4, 5, 6} as malignant. Distribution of density for each BI-RADS class is presented in

Table 11.2. In total, we used 3,272 MG images to conduct our experiments.

Table 11.2: Distribution of breast density in each BI-RADS class in the data-sets used in our study.

Purpose	**Data-set**	**BI-RADS**	**Class**				**Total**
			A	**B**	**C**	**D**	**All**
Training	**DDSM**	Benign	54	215	228	224	1,018
Training	**BCDR**	Benign	34	16	23	7	80
Training	**INbreast**	Benign	12	4	13	6	35
Training	**DDSM**	Malignant	219	693	513	291	1,716
Training	**BCDR**	Malignant	16	15	22	2	55
Training	**INbreast**	Malignant	30	32	8	2	72
Test	**UCHCDM**	Malignant	46	164	66	0	296

Pre-processing and Data-augmentation for Semantic Segmentation U-Net++ Model

The mammographic images are first pre-processed by employing normalization by scaling the image intensity to the range of [0, 255]. To segment mass(es) in the MG images, we generated different ground truth maps (GTMs) for the masses using the associated pixel-level boundary of the mass lesions given by the data-sets. All pixels in the GTM are labeled as belonging to the background (0) or breast lesion (255) classes. To deal with the small training data-set and avoiding overfitting our model, we applied data augmentation to the training MG images and it's corresponding GTMs (that contain mass(es) by image rotation by (-15,

15) degrees, translation up and down by (-10%, 10%), scaling in and out by 0.2, and left-right flips. All full MGs images and it's corresponding GTMs are re-sized to 512×512. The images that belong to the same patient are either in the training data-set. In Fig. 11.8, the upper row shows MG images for two patients, lesions are surrounded by rectangular GT in red color. The lower row shows the GT segmentations of the lesions for the same two patients shown in the upper row. (a) and (b), (c) and (d) images are examples for images extracted from the DDSM data-set and INbreast, respectively, for pre-training the models.

Whole Image Classifier ResNet50 Model

We proposed an end-to-end baseline architecture to classify whole MG images into normal or cancer. The network is pre-trained using patches extracted from a fully annotated data-set with ROI information normal, cancer calcification, and cancer mass, malignant mass. The patch classifier's weight parameters are then used to initialize the weight parameters of the whole image classifier model. After the network is pre-trained with patches, it is fine-tuned using data-sets without ROI annotations (normal vs. cancer).

We extracted balanced images for every class (normal, cancer calcification, cancer mass) from the DDSM, and BCDR mammography database to develop the patch and whole image classifier. We used the digital INbreast data-set to fine-tune the whole image classifier on two classes (normal, cancer). We evaluated the ResNet50 network for training the patch image classifiers to attain the best performance. We tested the proposed classifier for whole images on the current test images of the UCHCDM data-set. The DL structure for converting a patch classifier into a whole image classifier by changing the FC layers on top (see Fig. 11.7).

Training Strategy for Whole Image Classifier ResNet50 Model

There are two steps in training the ResNet50 model for a whole image classifier from scratch. The first step of training is to train a patch classification model. We use the DL models that have their weights pre-trained on the ImageNet database and we changed the top FC layers for three classes classification and

pre-train as follows:

- Set learning rate to $1e^{-3}$, freeze the top layers and train for 100 epochs using DDSM and BCDR patches.
- Set learning rate to $1e^{-5}$, unfreeze all layers and fine-tune for 200 epochs using INbreast patches.

Adam is used as an optimizer and the batch size is set to 16. The second step of converting the patch classifier into a whole image classifier is by altering the input image size from patch to whole image. The input image size of the pre-training patches are (512×512), however the input image size of whole images is 1024×1024. We proportionally increase the feature map size for every convolutional layer. We modified the top FC layers to classify 2 classes (normal, cancer). We used the same weights of the intermediate layers of the pre-trained patch classifiers. Similar to the patch network training, we used the following training strategy:

- Set learning rate to $1e^{-3}$, freeze the top layers and train for 50 epochs using the DDSM and BCDR whole images.
- Set learning rate to $1e^{-5}$, unfreeze all layers and fine-tune for 30 epochs using the INbreast whole images.

We used Adam to optimize the loss function, and used a small batch size of 4 for whole image training. We used data augmentation for both patch and whole image training. We assign all images with BI-RADS readings of 1 as normal images; 2: 6 as cancerous images.

11.3.3 Evaluation Metrics

We used the following performance validation metrics to evaluate their method: sensitivity (Sen.), specificity (Spe.), accuracy (Acc.), ROC curve, and F1 metric.

The sensitivity (SEN), recall, hit rate, or true positive rate (TPR) represents the proportion of positive cases

correctly classified (Equation 11.5). Where TP is true positive, FN is a false negative, TN is a true negative, and FP is false positive.

$$Sen = \frac{TP}{TP + FN}, \tag{11.5}$$

Precision (also called positive predictive value (PPV)) is the fraction of positive cases among the retrieved instances (Equation 11.6).

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{11.6}$$

The specificity (Spe.), selectivity or true negative rate (TNR) is the proportion of actual negative cases which are correctly classified (Equation 11.7).

$$Spe = \frac{TN}{TN + FP}, \tag{11.7}$$

The accuracy (ACC) represents the proportion of true results (both true positives and true negatives) (Equation 11.8).

$$ACC = \frac{TP + TN}{TN + TP + FN + FP}, \tag{11.8}$$

The Receiver operating characteristic (ROC) curve is the plot between sensitivity and (1- specificity). (1-specificity) is also known as false positive rate (FPR).

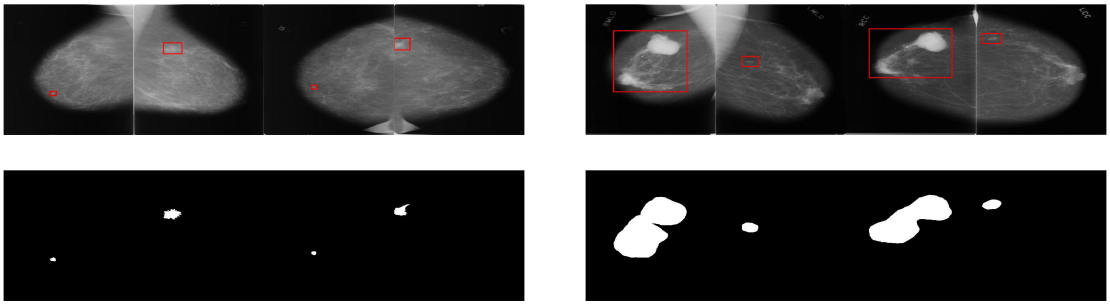
The AUC measure is computed just by obtaining the area under of ROC curve (Equation 11.9).

$$AUC = \frac{1 + TPR - FPR}{2}, \tag{11.9}$$

The F1 score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0, (Equation 11.10).

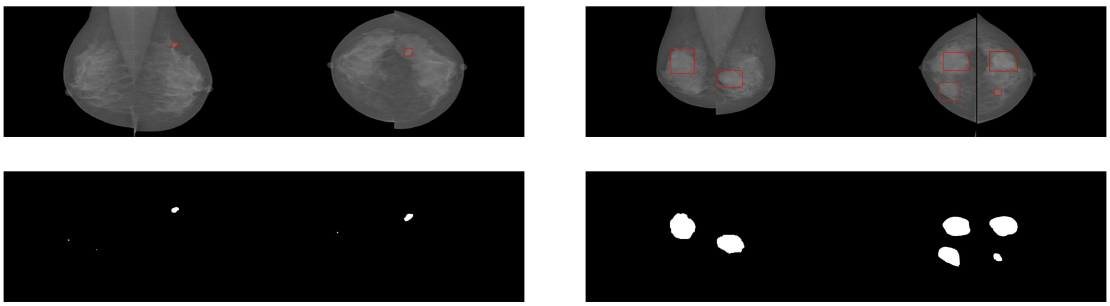
$$F1_{score} = \frac{2 \times precision \times recall}{precision + recall}, \tag{11.10}$$

The F1-score combines the trade-offs of precision and recall and outputs a single number reflecting the goodness of a classifier in the presence of imbalanced data-sets.



(a) Patient has scattered areas of fibroglandular. The left breast has malignant (BI-RADS 4) mass of irregular AD shape and spiculated margins, and the right breast has benign (BI-RADS 2) mass of round shape and circumscribed margins.

(b) Patient breast is almost entirely fatty. The left breast has benign (BI-RADS 2) mass of oval shape and circumscribed margins, and the right breast has malignant (BI-RADS 5) mass of lobulated shape and circumscribed margins.



(c) Patient has heterogeneously breast density. The left breast has benign (BI-RADS 3) mass(es) of oval shape and circumscribed margins. The right breast has small number of diffused micros calcifications.

(d) Patient has heterogeneously breast density. The right and left breast has benign (BI-RADS 2) mass(es) of oval shape and circumscribed margins. One of the mass(es) are not seen in the MLO right and left breast.

Fig. 11.8: The upper row shows MG images for two patients, lesions are surrounded by rectangular GT in red color. The lower row shows the GT segmentations of the lesions for the same two patients shown in the upper row. (a) and (b), (c) and (d) images are examples for images extracted from the DDSM data-set and INbreast, respectively, for pre-training the models.

11.4 Results

We have presented a training strategy for performing one-shot recognition by the deep CNN Siamese model. We compared the performance of the proposed Siamese model to the performance of the state-of-the-arts base-line models, U-Net++ and ResNet50 (see Table 11.3). The deep CNN Siamese model is tested on pairs of MG images (current and prior). While the U-Net++ and the ResNet50 are tested on current MG images. To test the U-Net++ model ability to detect and accurately localize lesions, we evaluated the predictions on the UCHCDM data-set using the ROC curve. The ROC curve shows sensitivity (fraction of correctly localized lesions) as a function of the number of false-positive (see Fig 11.10). A detection was considered correct if the IOU of the segmented lesion with the ground truth maps is 0.3. The proposed Siamese model has a sensitivity of 0.932 and FPR of 0.068. The U-Net++ model has a sensitivity of 0.885 and an FPR of 0.116 (see Fig. 11.10). The single network classifier has a sensitivity of 0.816 and an FPR of 0.184.

Figures 11.11, 11.12, and 11.13 show the detection of the U-Net++ model in detecting tumors in the UCHCDM data-set. Every pair of images show different views for the same breast (CC and MLO). Prior images in Figs. 11.11, 11.12, and 11.13 are detected as normal, while current images are detected as cancerous images. The last two columns show the detection of the U-Net model for prior and current year images for the same patient from two different views. In Figs. 11.15, 11.16, 11.17, 11.18, 11.19, and 11.20, the U-Net++ model detect lesions in prior year images that is small in size comparing to current year images. The GT images of these detected lesions are normal. The U-Net++ was able to detect small lesions that developed in years to become cancerous tumors.

Table 11.3: Comparison of the best one-shot accuracy of Siamese-Net against baseline models.

Model	Acc.	Sen.	Spe.	Precision	FPR	F1_score	AUC
Siamese-Net	0.932	0.932	0.932	0.935	0.068	0.932	0.992
U-Net++	0.884	0.885	0.885	0.841	0.116	0.884	0.968
ResNet50	0.816	0.816	0.816	0.860	0.184	0.816	0.852

11.5 Discussion and Conclusions

We have proposed a Siamese CNN approach, which achieved an AUC of 0.992 on the test UCHCDM dataset. We also compared the performance of the Siamese CNN model with a baseline CNN model. The accuracy of the classification of MG images increased by 14.22% when prior images were used to train the DL CNN model. Using a baseline CNN model, the accuracy dropped from 0.932 to 0.816. Results show that using prior-year images in training DL models is a promising approach.

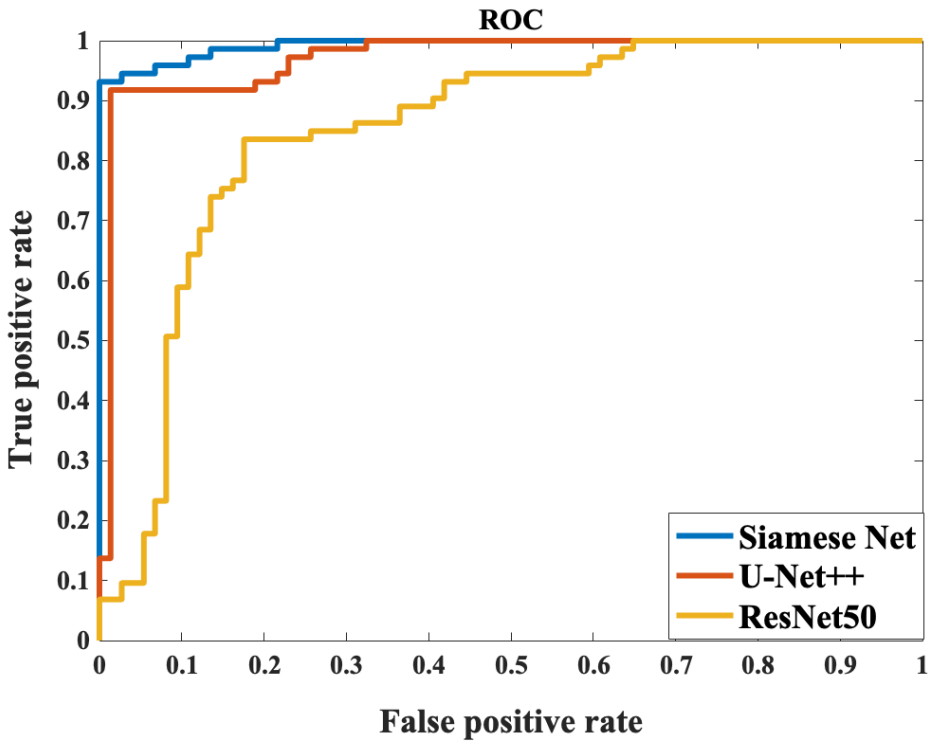


Fig. 11.10: Classification performance.

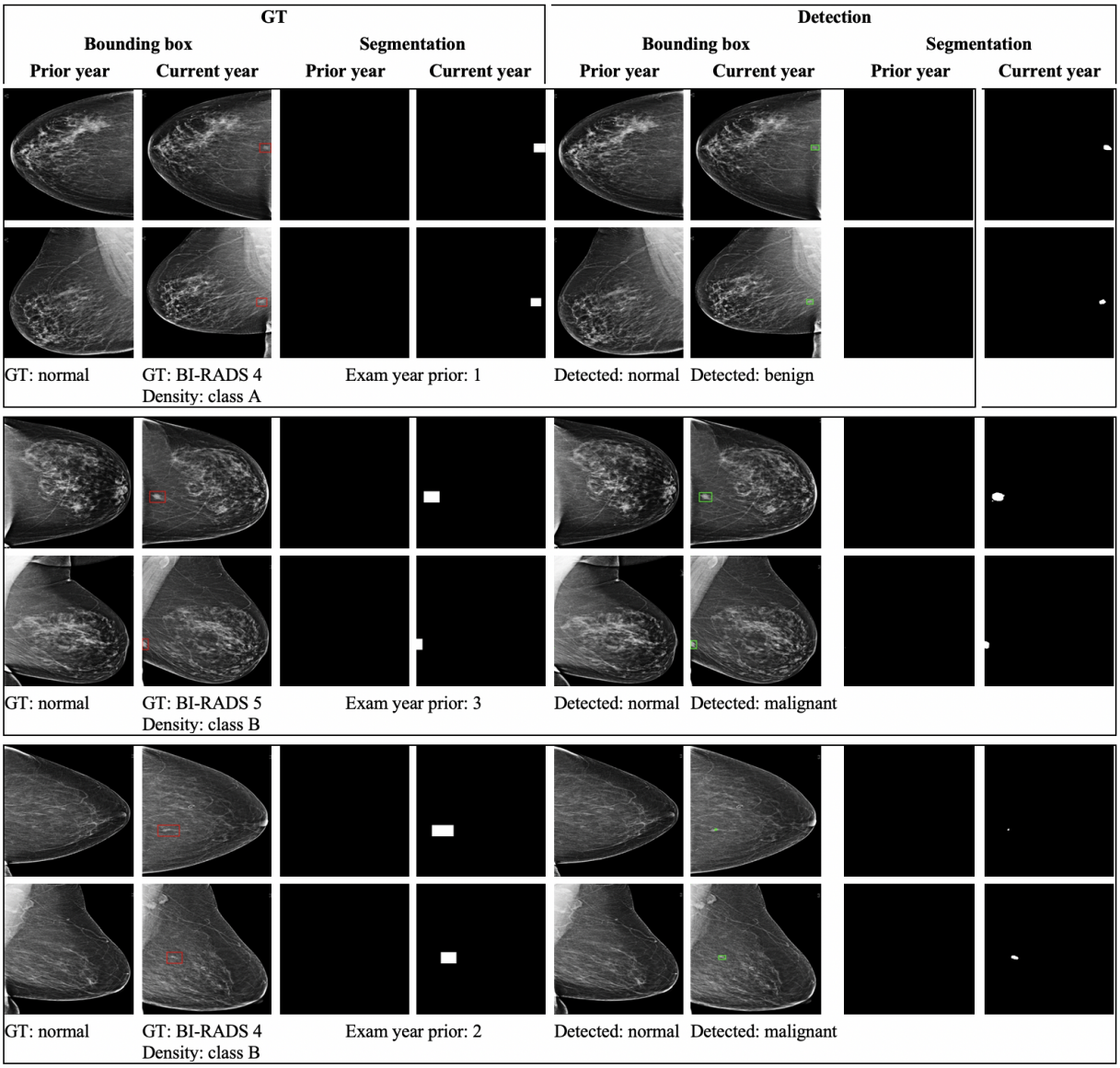


Fig. 11.11: Detection of tumors in current-year and prior-year images.

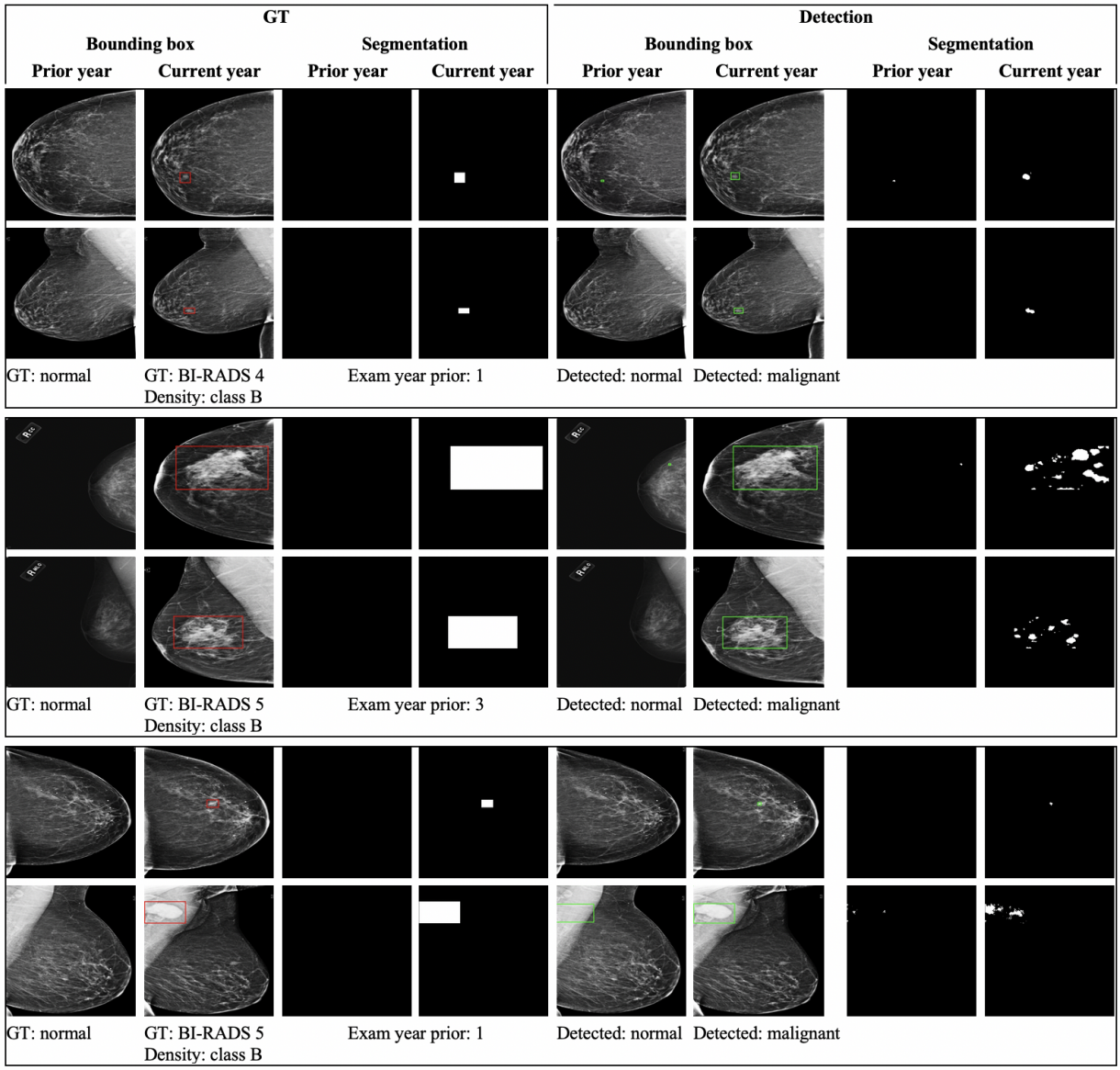


Fig. 11.12: Detection of tumors in current-year and prior-year images.

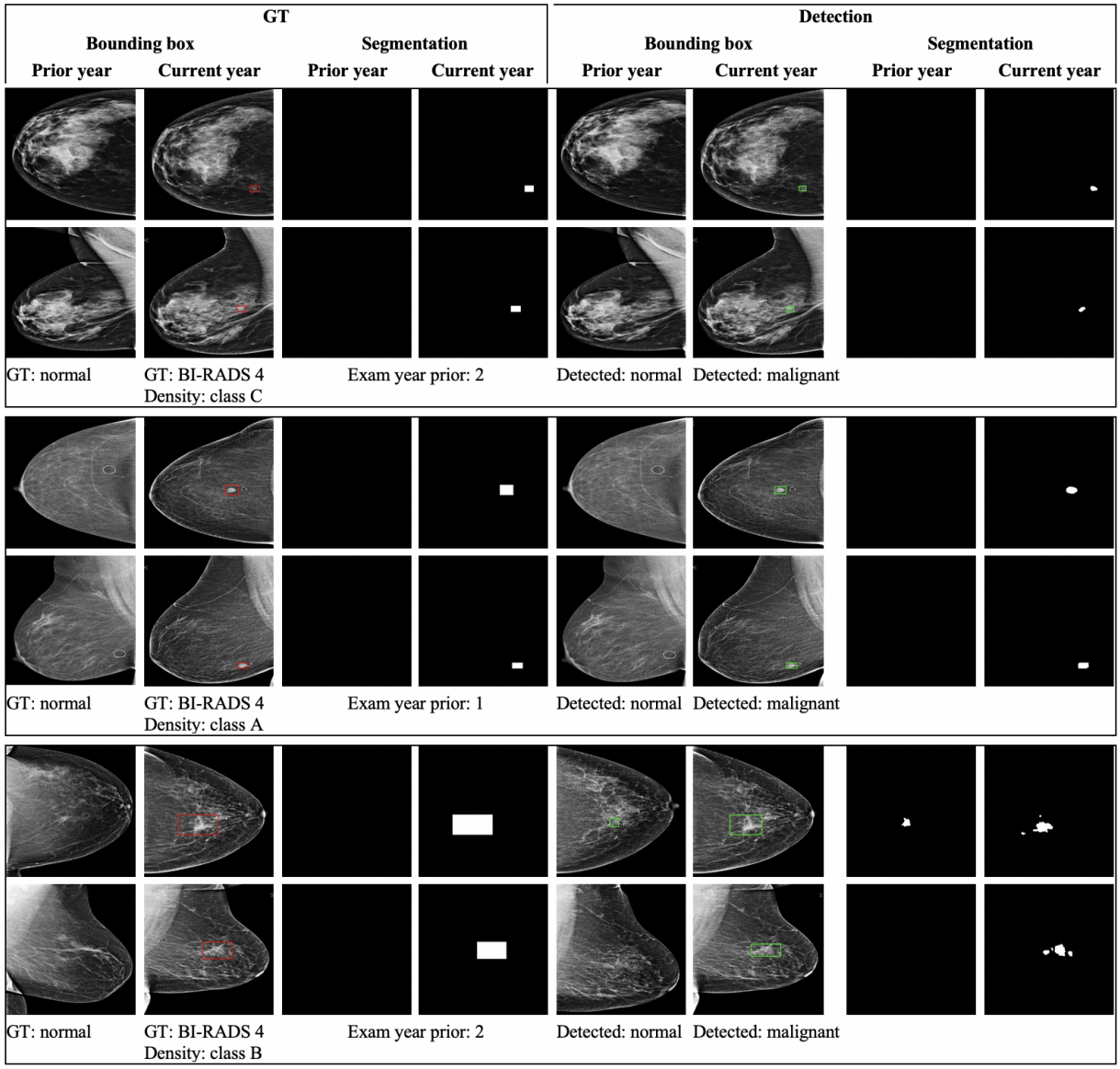


Fig. 11.13: Detection of tumors in current-year and prior-year images.

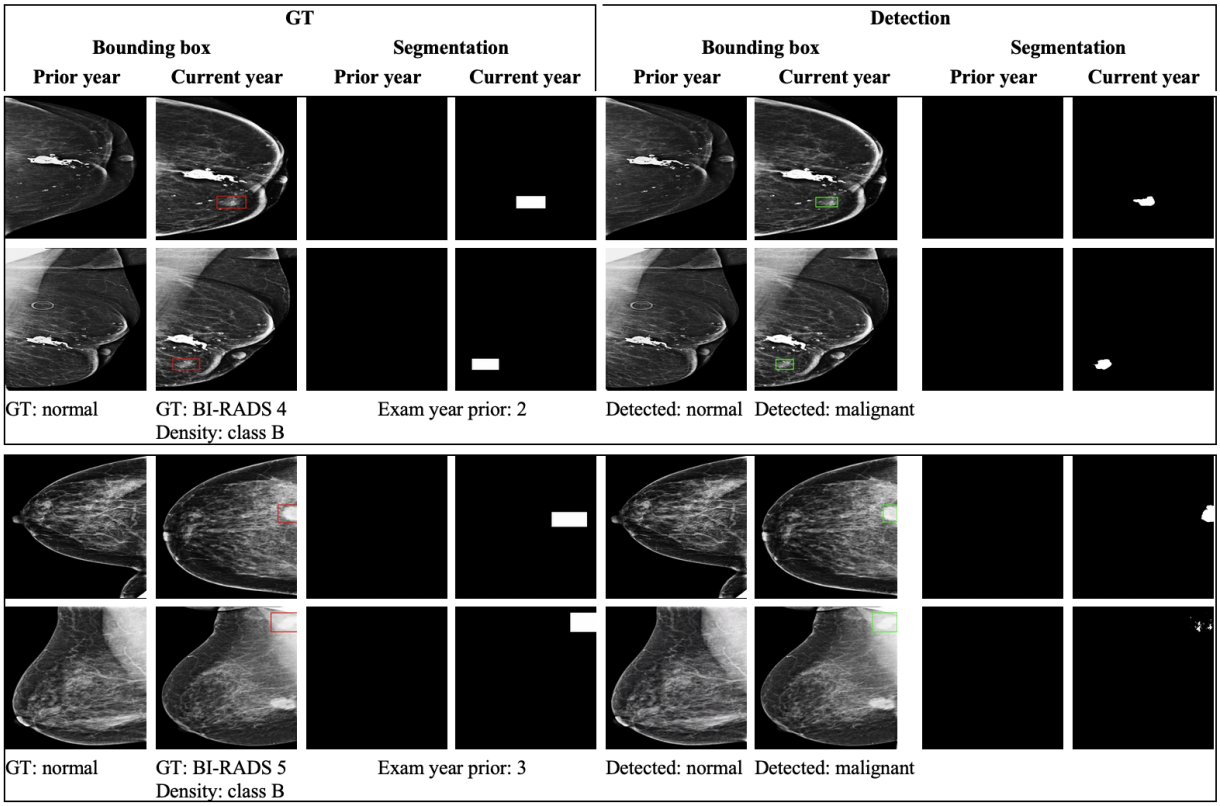


Fig. 11.14: Detection of tumors in current-year and prior-year images.

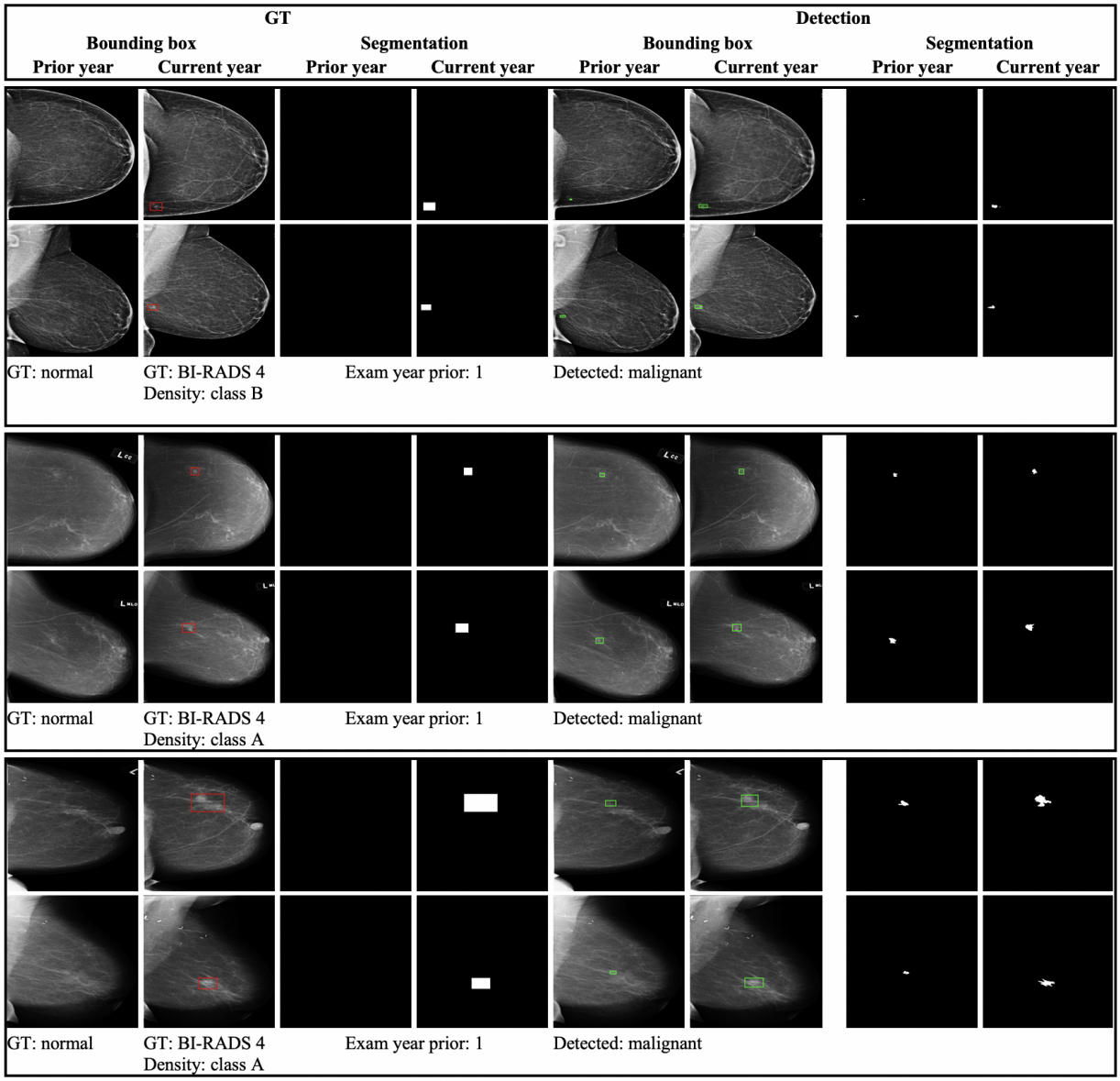


Fig. 11.15: Detection of tumors in current-year and prior-year images.

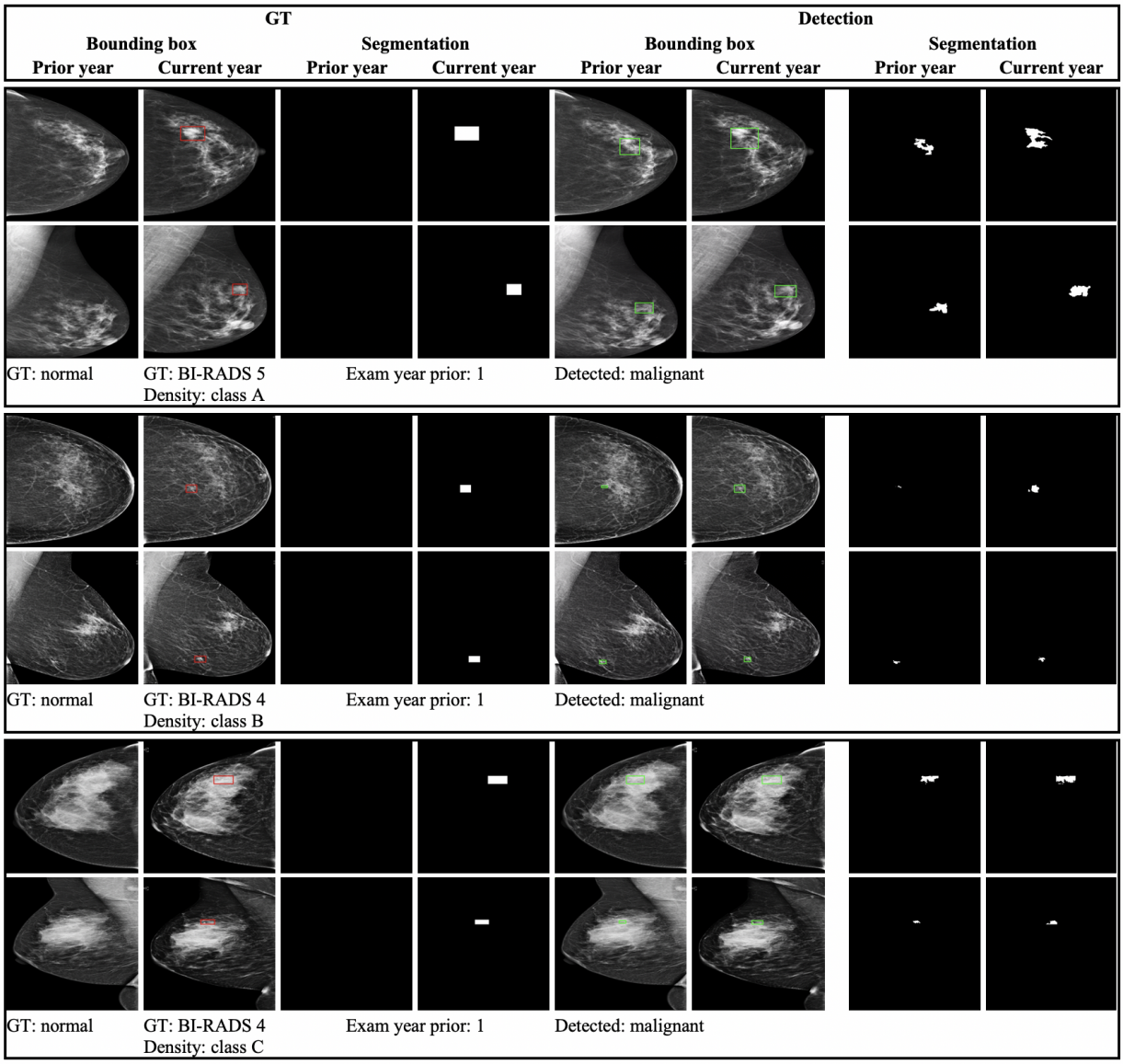


Fig. 11.16: Detection of tumors in current-year and prior-year images.

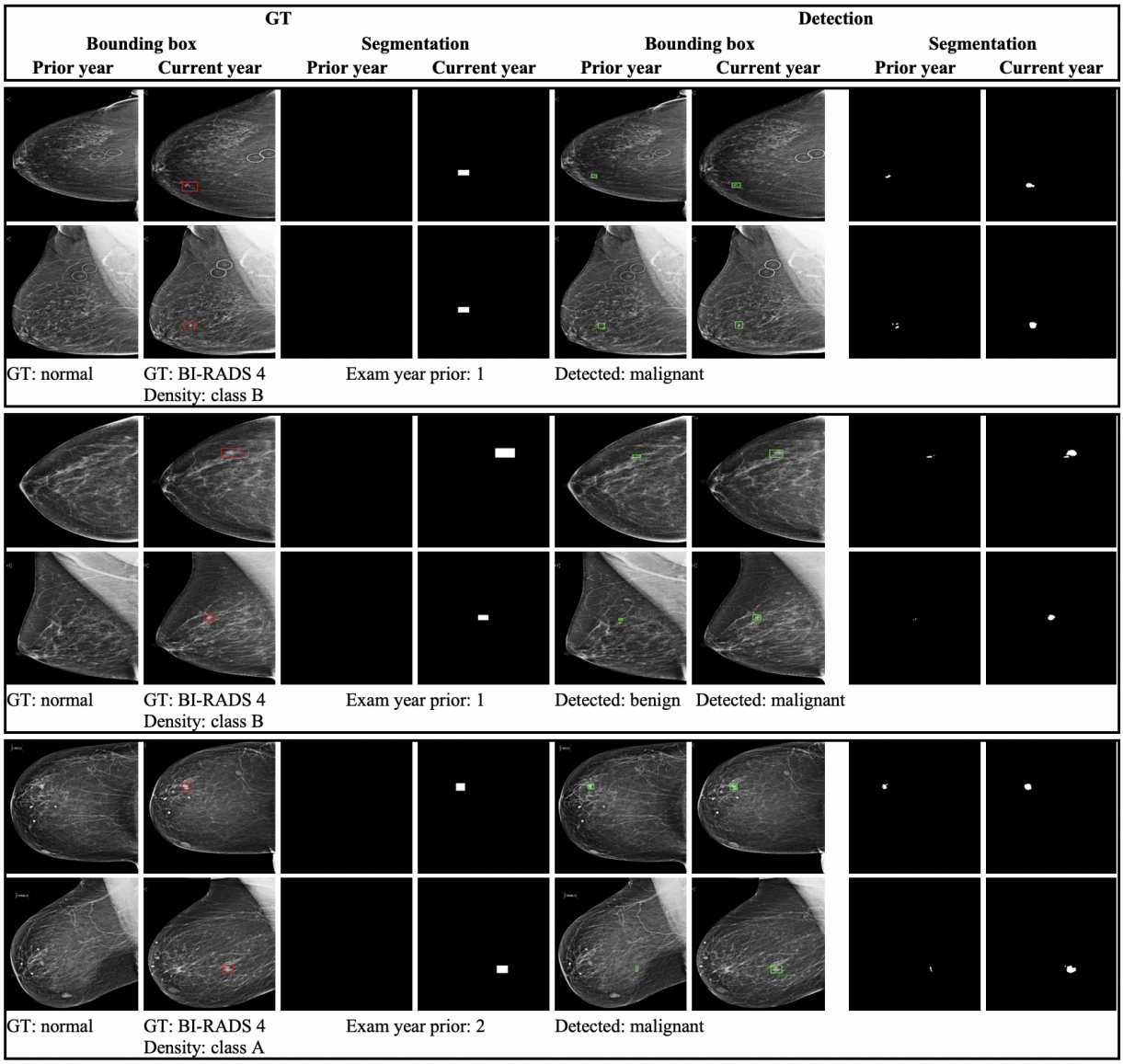


Fig. 11.17: Detection of tumors in current-year and prior-year images.

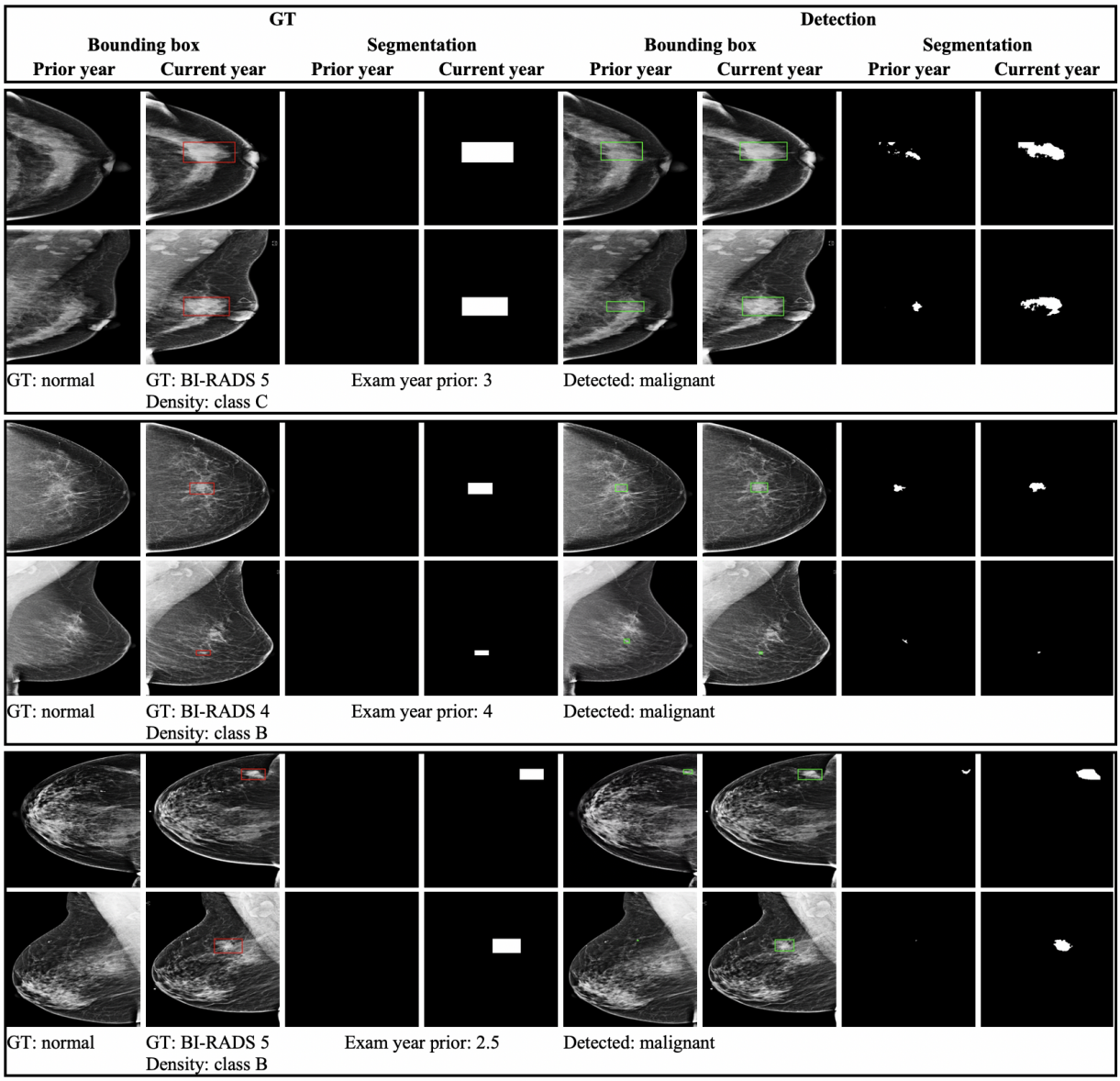


Fig. 11.18: Detection of tumors in current-year and prior-year images.

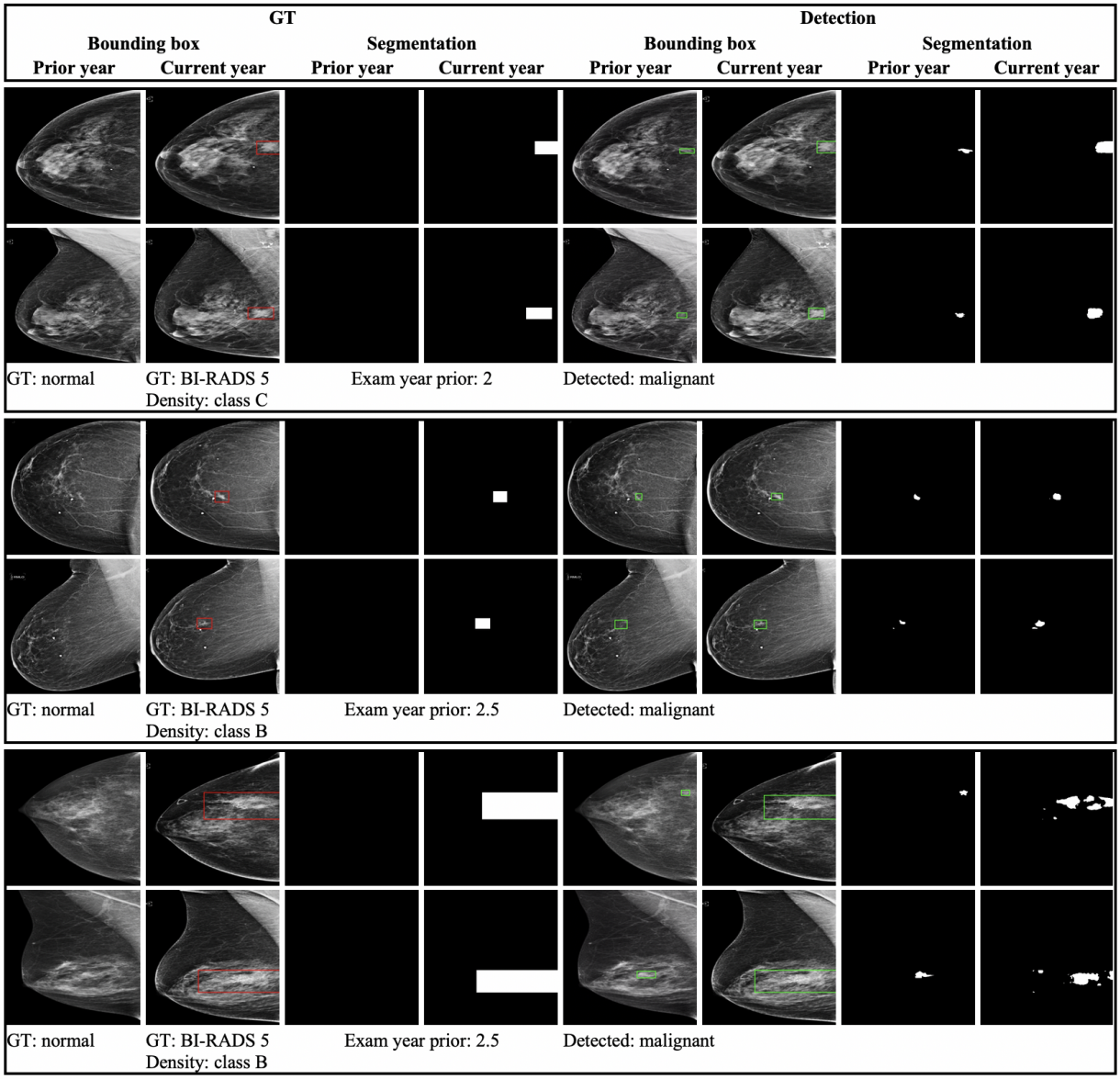


Fig. 11.19: Detection of tumors in current-year and prior-year images.

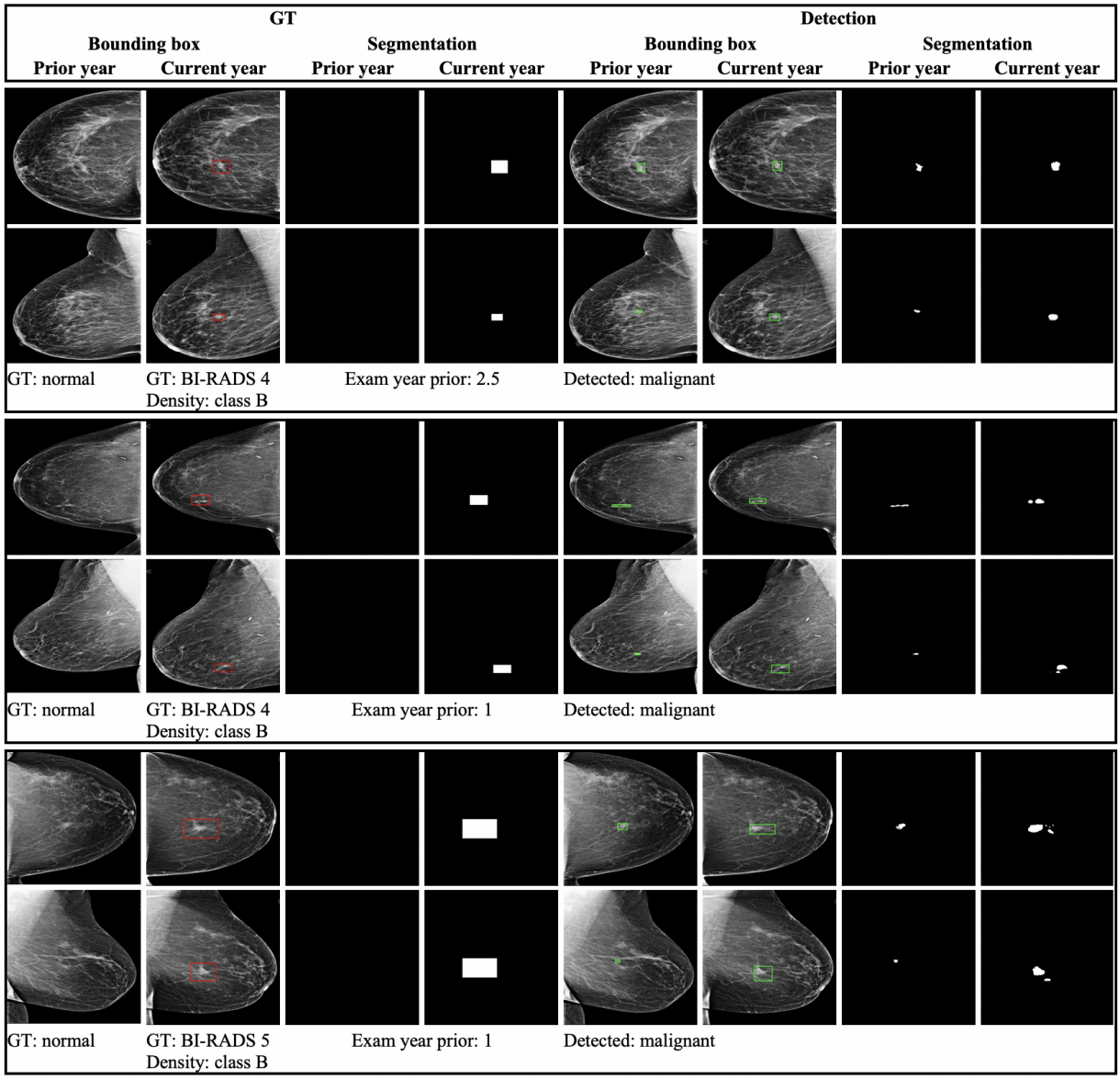


Fig. 11.20: Detection of tumors in current-year and prior-year images.

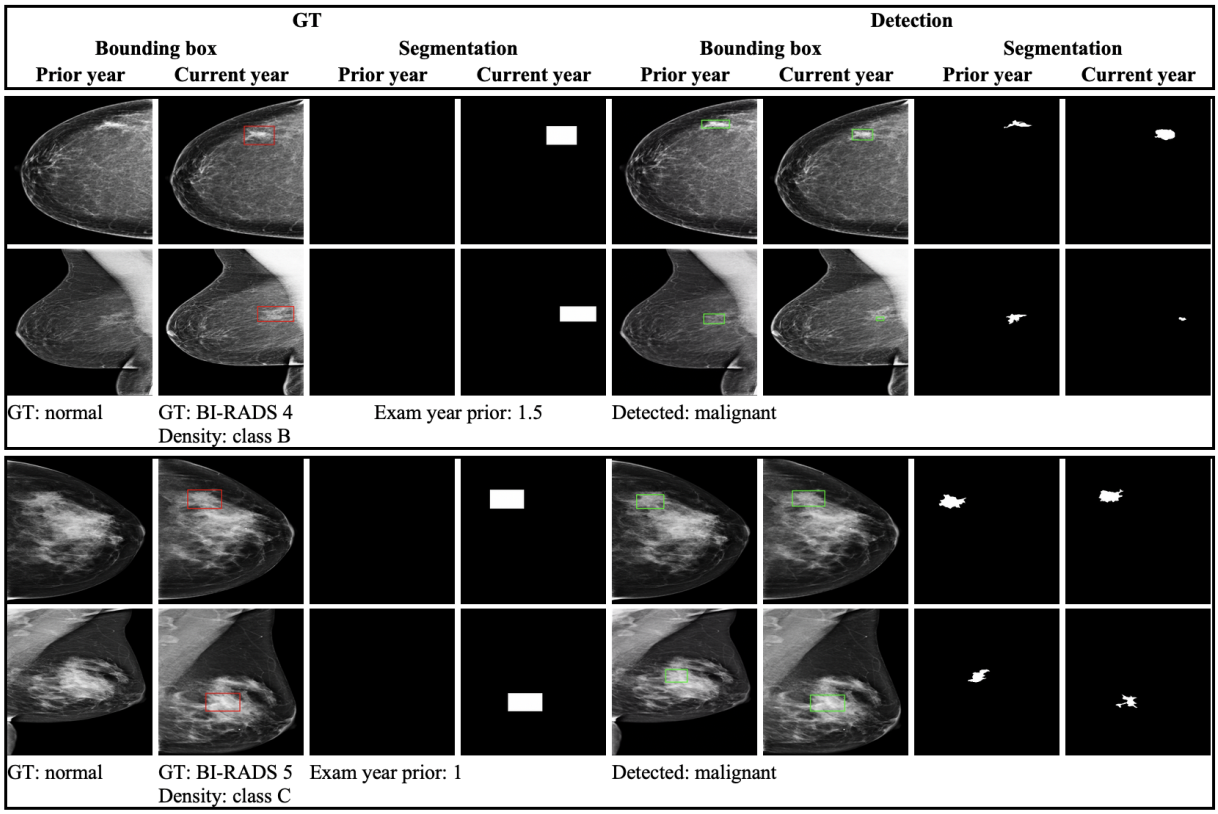


Fig. 11.21: Detection of tumors in current-year and prior-year images.

Chapter 12

Future Work: Generative Adversarial Network for Data Augmentation of Breast Masses in Mammogram Images

12.1 Abstract

Deep learning (DL) approaches, specially convolution neural networks (CNN), have recently shown promising results in detecting, classification of breast tumors in mammograms (MGs). However, such models are constrained by the limited size of publicly available mammography data-sets such as DDSM, INbreast, BCDR. Several research studies have been conducted using private mammography data-sets, which are not publicly available for privacy concerns, which arise limitation on the reproducibility of their work. Moreover, the size of normal images is outnumbering those cancer images with tumor findings. This leads us to unbalanced class problems in training DL models.

Recent studies have used synthetically MG to enlarge their data-set to avoid over-training their model and avoid biased classifier, which can be done using simple data augmentation techniques (e.g. horizontal flipping, zoom in and out, rotations). In the last chapter, we have developed a Siamese network trained on prior MG images (normal) and current year MG images (normal or cancerous). The main problem that faced us is none of the publicly available databases have prior and current images for the same patient except for few cases in the BCDR data-set. Increasing the data-set with more cases of prior-current images would improve the performance of the proposed model. We propose a generative adversarial network (GAN) that

can be used to synthesize lesions in MG images, thus we utilize to create more images for the prior images but with different implanted cancerous lesions.

12.2 Background

Singh et al [202] presented a conditional GAN model (cGAN) to segment mammographic lesions from an ROI. The proposed generative model learns the lesion representations to create binary masks and the adversarial network learns features that discriminate the real masses from the generated binary masks. GANs have been used for data augmentation in several recent works [20, 85, 193]. Shams et al [193] proposed a GAN-based mammogram classification method to deal with data scarcity and limited availability of annotated data. Al-Dhabyani et al. [20] used GAN to augment ultrasound images and boost the performance of their proposed classifier. Similarly, Guan et al. [85] proposed to use the GAN model to generate synthetically patches that have regions of interest (ROIs) that are normal or abnormal (cancer/tumor). In this work, we first show that GANs are capable of generating high-resolution synthetic MG images and the implant lesions are left for future work.

12.3 Material and Methods

12.3.1 Data-sets

We conducted our initial experiment on the publicity available INbreast fully field digital (FFDM) data-set. The INbreast data-set contains 115 patients and 410 mammograms including both CC and MLO views. We analyzed each view separately. The INbreast database includes radiologists' BI-RADS assessment categories which are defined as follows: incomplete exam: 0, normal: 1, benign: 2, probably benign: 3, suspicious: 4, highly suggestive of malignancy: 5, and known biopsy-proven cancer: 6. We assigned all images with BI-RADS 1 as negative and images with BI-RADS 2:6 as cancerous.

12.3.2 Proposed GAN Model

A generative adversarial network (GAN) is a type of DL network that can generate data with similar characteristics as the input training data. A GAN consists of two networks, discriminator D(X) and generator G(Z), that train together. Given a vector of random values Z as input, the generator network generates data with the same structure as the training data X. Given batches of data containing observations from both the training data and generated data from the generator, the discriminator network attempts to classify the observations as ”real” or ”generated”, as shown in Fig. 12.1. The generator model takes uniform 100-dimensional noise distribution Z as input and the result is a four-dimensional tensor that is used as the beginning of the convolution stack. Thus, Z is projected into a small spatial convolutional representation with several feature maps. The generator consists of five transpose convolution layers (conv.) followed by batch normalization (BN) layer, then a ReLU activation function, except for the last convolutional layer is feed to Sigmoid output. For discriminator, it has five Conv. layers followed by the BN layer, ReLU layer, and dropout layer of 25%, the last Conv. layer is flattened and fed into sigmoid output. Weights are initialized from a zero-centered normal distribution. For a detailed architecture, refer to Fig. 12.2.

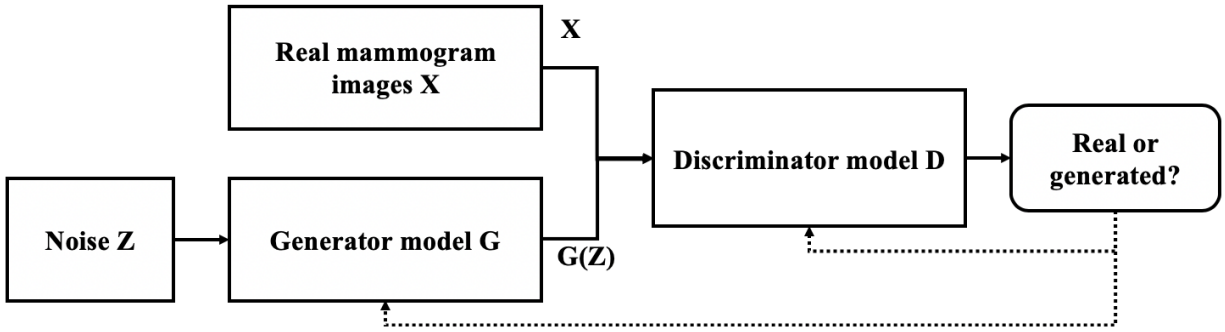
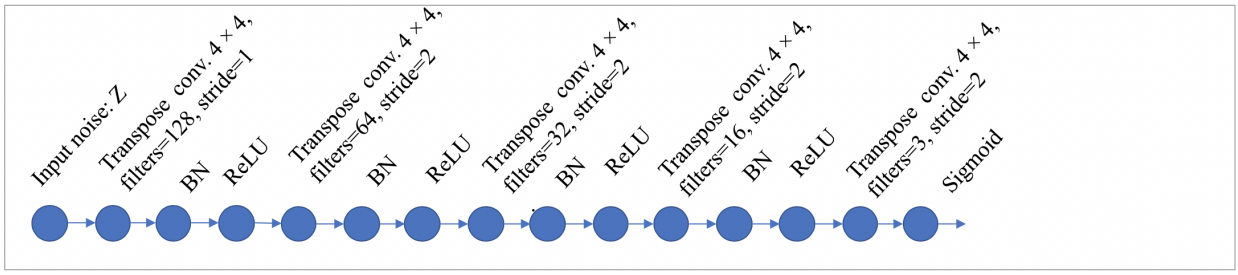


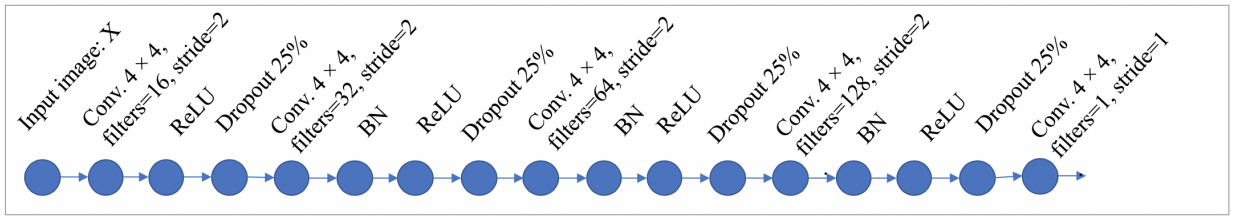
Fig. 12.1: Generative adversarial network framework.

To train the proposed GAN, we train both networks simultaneously to maximize the performance of both.

The objective of the generator is to generate data that the discriminator classifies as ”real”. To maximize the probability that images from the generator are classified as real by the discriminator, minimize the



(a) Generator model.



(b) Discriminator model.

Fig. 12.2: Architecture of the proposed generative adversarial network.

negative log-likelihood function. The loss function for the generator is given by:

$$loss_{Generator} = -mean(log(\sigma(\hat{Y}_{Generated}))), \tag{12.1}$$

where σ denotes the sigmoid function, and $\hat{Y}_{Generated}$ denotes the output of the discriminator with generated data input.

The objective of the discriminator is to not be "fooled" by the generator. To maximize the probability that the discriminator successfully discriminates between the real and generated images, minimize the sum of the corresponding negative log likelihood functions. The output of the discriminator corresponds to the probabilities the input belongs to the "real" class. For the generated data, we use the probabilities corresponding to the "generated" class $1 - \sigma(\hat{Y}_{Generated})$.

$$loss_{Discriminator} = -mean(log(0.9 \times \sigma(\hat{Y}_{Real}))) - mean(log(1 - \sigma(\hat{Y}_{Generated}))), \tag{12.2}$$

where $\hat{Y}_{Real}$ denotes the output of the discriminator with real data input.

We train our GAN model with a mini-batch size of 32 for 5000 epochs. Due to the nature of mammogram data-sets, we could only implement a limited range of transformations. In particular, we applied a horizontal flip to generate more images. We use ADAM as our optimization function.

The learning rate of the generator is set to 0.0002 and the learning rate of the discriminator to 0.0001. For both networks, we use a gradient decay factor of 0.5 and a squared gradient decay factor of 0.999. Train the model using a custom training loop. Loop over the training data and update the network parameters at each iteration.

To monitor the training progress, display a batch of generated images using a held-out array of random values to input into the generator. For each epoch, we shuffle the data-set and loop over mini-batches of data. For each mini-batch, we normalize the data so that the pixels take values in the range [-1, 1]. We use a drop-out of 0.25 in the discriminator network. After every 100 iterations, we display a batch of generated images for fixed held-out generator input. To monitor training progress, we used a held-out batch of fixed 64 1-by-1-by-100 arrays of random values to input into the generator. The size of the input image will be modified in the future to 1024×1024 to generate high-resolution images. We used Matlab/2019b for the implementation of our network and the model was trained on Tesla V100-PCIE-16GB GPU.

12.4 Initial Results

This section contains initial results to show that GANs are capable of generating synthetic mammogram images. Figure 12.3 shows the first iterations of the training models where the input is noise. Figures 12.4 and 12.5 show the different iterations of the training models where the network starts to generate synthetic MG images. Figure 12.6 shows the outputs of the GAN model.

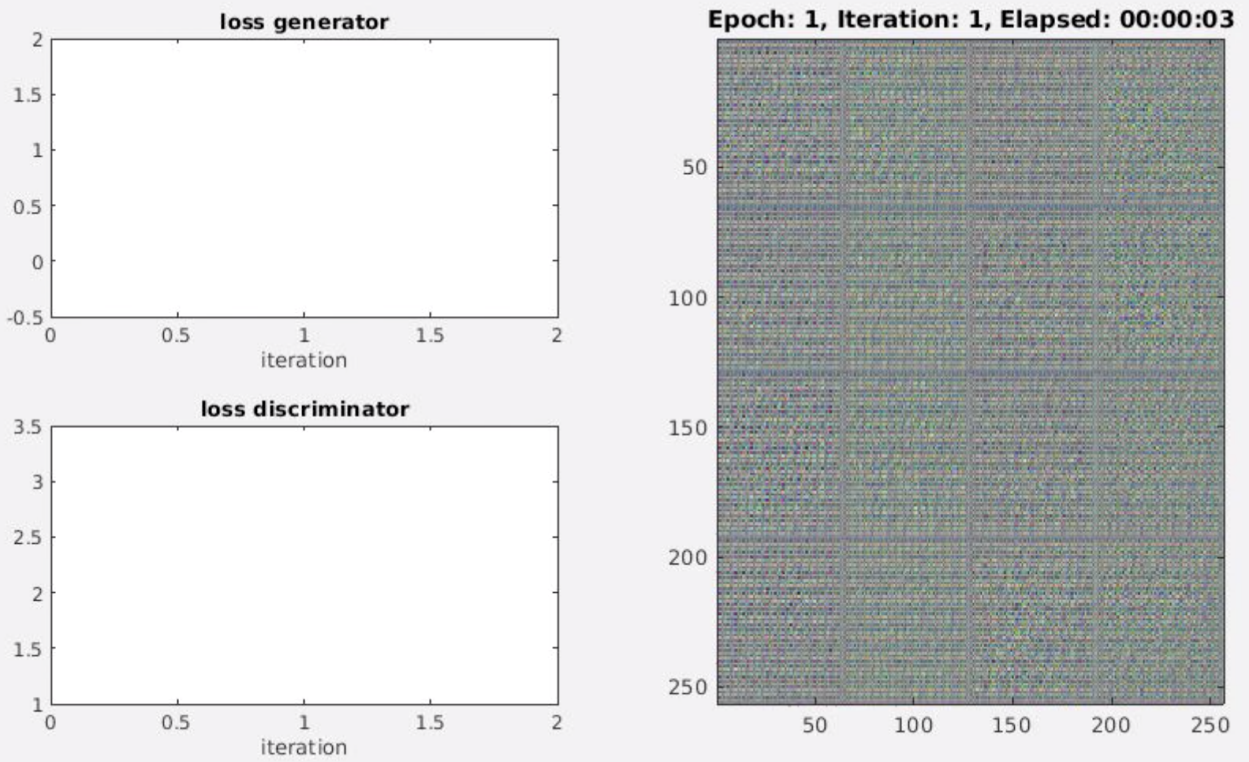


Fig. 12.3: Very first iterations using GAN.

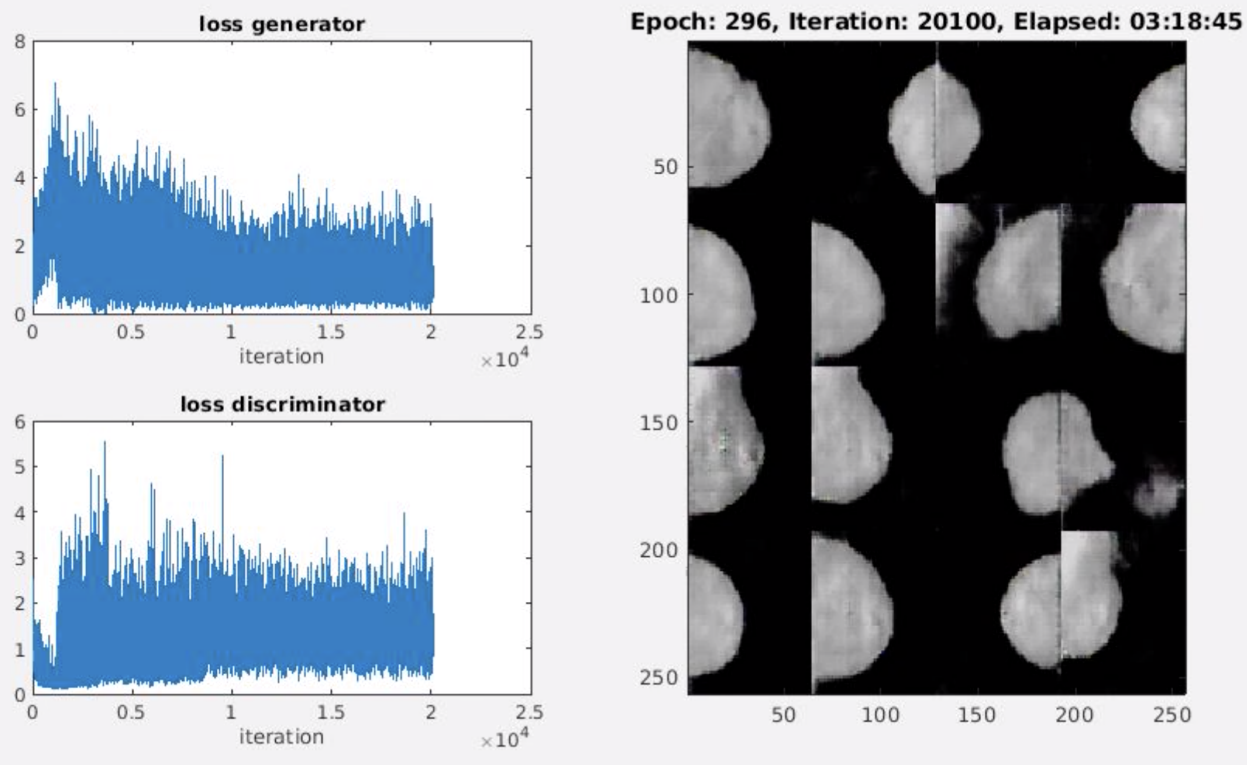


Fig. 12.4: Different iterations of the training models where the network start generate synthetic MG images.

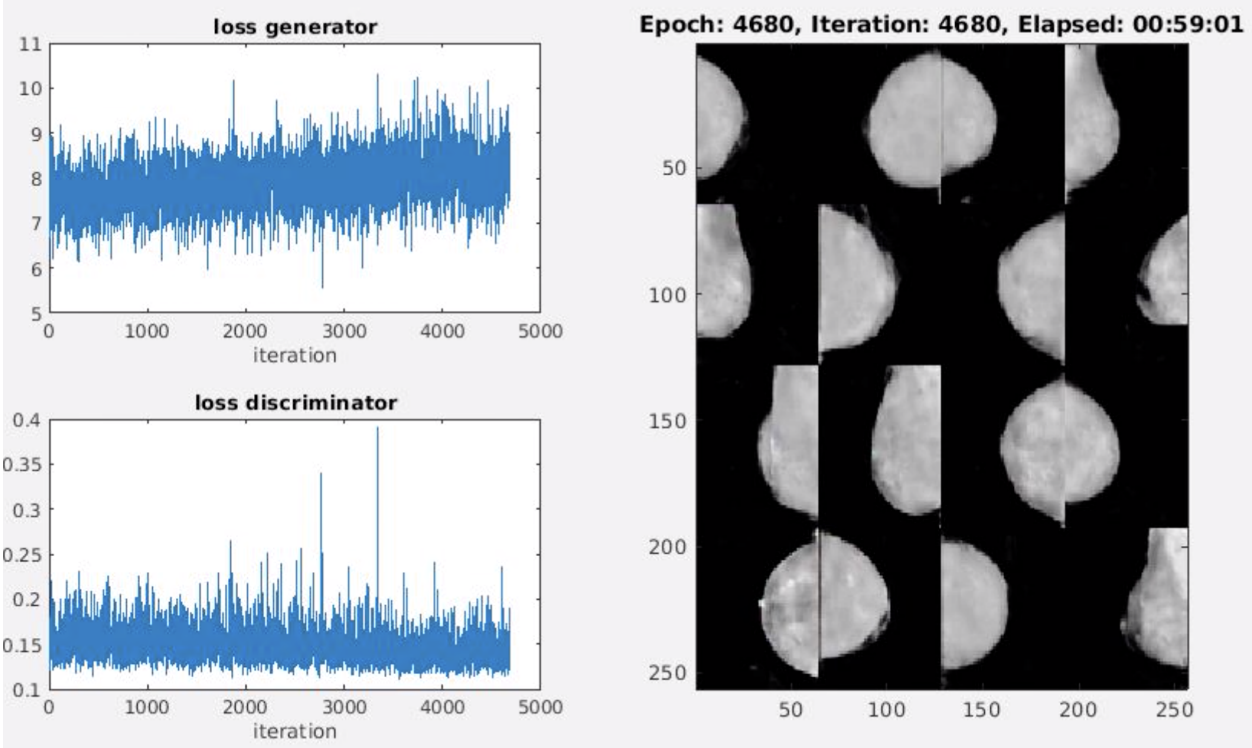


Fig. 12.5: Different iterations of the training models where the network start generate synthetic MG images.

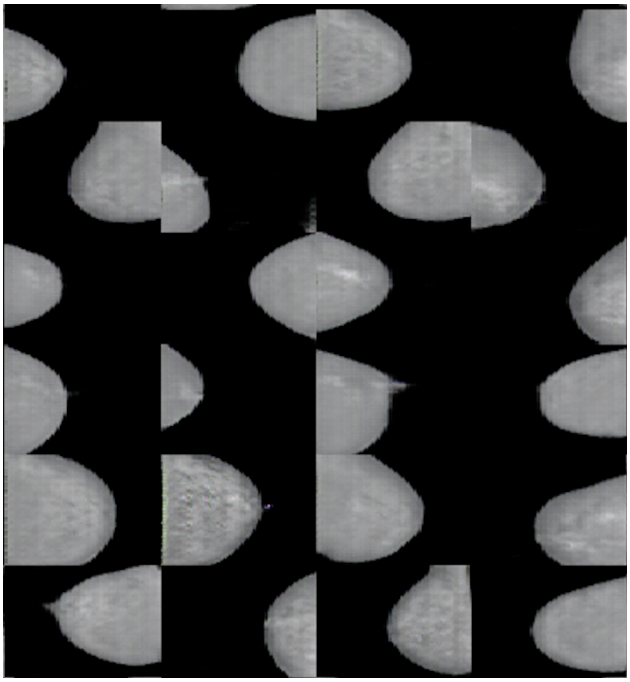


Fig. 12.6: Outputs of the GAN model.

12.5 Discussion

Initial results using the GAN model show promising results for generating synthetic MG images. The next step is to generate high-resolution high images and generate synthetic MG images that have synthetic tumors in it.

Chapter 13

Discussion and Conclusions

In this dissertation, we summarize 83 research studies for applying CNNs on various tasks in mammography. This extended survey focuses on finding the best practices used in these research studies to improve diagnosis accuracy. Also, it provides a deep insight into the architecture of CNNs used for various tasks. Furthermore, it describes the most common publicly available MG repositories and highlights their main features and strengths (Chapters 5 and 6).

We proposed several deep learning models for the precise segmentation of mass lesions in mammograms. In Chapter 8, we proposed Vanilla U-Net model. The intersection over union (IOU) of detecting mass lesions of the proposed Vanilla U-Net model improves from 91.7% to 93.4% when training with the augmented dataset. The proposed Vanilla U-Net model yields an improvement of 7.46% in mean Bf-score relative to that of the original U-Net model.

In Chapter 9, we proposed RAttU-Net model. The proposed RAttU-Net model yields an improvement of 4.24%, 30.02% in the DI and 2.93%, 13.39% in the IOU, respectively, relative to that of the Vanilla U-Net model and the original U-Net. The proposed RAttU-Net model can detect lesions in the heterogeneously dense test images with a mean accuracy of 0.995 and a mean IOU of 0.930. Moreover, the proposed RAttU-Net model can detect lesions in the extreme dense test images with a mean accuracy of 0.94 and mean IOU of 0.928. The proposed RAttU-Net model yields an improvement of 4.24%, 30.02% in the DI and 2.93%, 13.39% in the IOU, respectively, relative to that of the vanilla U-Net model and the original U-Net.

In Chapter 10, we introduced the density-based attention networks to achieve better detection of lesions in mammograms especially in the heterogeneous dense and dense images. The RAttU-Net (pre-trained on density masks) is superior in detecting masses in heterogeneous and dense MGs with a high mean IOU of 0.912 and 0.858, respectively than other models. AttU-Net (pre-trained on density masks) with high IOU of 0.858 and 0.833, respectively. The RAttU-Net (pre-trained on density masks) and U-Net++ (pre-trained on density masks) show near mean IOU. The The RAttU-Net show improvement in detection of tumors with IOU increase from 0.923 to 0.950 (fatty class), from 0.887 to 0.928 (scattered class), from 0.805 to 0.912 (heterogeneous class), and 0.757 to 0.858 (dense class).

In Chapter 11, we introduced the Siemens one-shot model to integrate prior-year images with current-year images to reduce false positives. The accuracy of the classification of MG images increased by 14.22% when prior images were used to train the DL CNN model. Using prior year images, FPR decreased from 0.852 to 0.068 and the accuracy of classification increase from 0.816 to 0.932.

In conclusion, in this dissertation, we introduced several deep learning models to detect and localize precise segments of lesions in whole MG images. Our proposed models find lesions in the challenging breast density cases (heterogeneously dense and extremely dense MG image). We used the One-shot model to integrate prior-year MG images in the training of the proposed models to reduce the false positives.

Bibliography

[1] Cbis-ddsm. https://mcl.nci.nih.gov/science-data/cbis-ddsm-1, last accessed: 3/2/2019.

[2] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[3] Q. Abbas. Deepcad: A computer-aided diagnosis system for mammographic masses using deep invariant features. *Computers*, 5(4):28, 2016.

[4] D. Abdelhafiz, S. Nabavi, R. Ammar, and C. Yang. Survey on deep convolutional neural networks in mammography. In *Computational Advances in Bio and Medical Sciences (ICCABS), 2017 IEEE 7th International Conference on*, pages 1–1. IEEE, 2017.

[5] D. Abdelhafiz, S. Nabavi, R. Ammar, and C. Yang. The effect of pre-processing on breast cancer detection using convolutional neural networks. *Poster session presented at the meeting of the IEEE International Symposium on Biomedical Imaging,Washington, DC*, 2018.

[6] D. Abdelhafiz, S. Nabavi, R. Ammar, C. Yang, and J. Bi. Convolutional neural network for automated mass segmentation in mammography. In *2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, pages 1–1. IEEE, 2018.

[7] D. Abdelhafiz, S. Nabavi, R. Ammar, C. Yang, and J. Bi. Residual deep learning system for mass segmentation and classification in mammography. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 475–484, 2019.

[8] D. Abdelhafiz, C. Yang, R. Ammar, and S. Nabavi. Additional file 1 of deep convolutional neural networks for mammography: advances, challenges and applications. *BMC Bioinformatics*, 20(11):281, Jun 2019.

[9] D. Abdelhafiz, C. Yang, R. Ammar, and S. Nabavi. Deep convolutional neural networks for mammography: advances, challenges and applications. *BMC Bioinformatics*, 20(11):281, Jun 2019.

[10] D. Abdelhafiz, C. Yang, R. Ammar, and S. Nabavi. Deep convolutional neural networks for mammography: advances, challenges and applications. *BMC bioinformatics*, 20(11):281, jun 2019.

[11] D. Abdelhafiz, C. Yang, R. Ammar, and S. Nabavi. Residual deep learning system for mass segmentation and classification in mammography. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. ACM-BCB '19, 2019.

[12] V. Agarwal and C. Carson. Using deep convolutional neural networks to predict semantic features of lesions in mammograms. *C231n Course Project Reports*, 2015.

[13] P. Agrawal, M. Vatsa, and R. Singh. Saliency based mass detection from screening mammograms. *Signal Processing*, 99:29–47, 2014.

[14] J. Ahmad, M. Sajjad, I. Mehmood, and S. W. Baik. Sinc: Saliency-injected neural codes for representation and efficient retrieval of medical radiographs. *PloS one*, 12(8):e0181707, 2017.

[15] C. K. Ahn, C. Heo, H. Jin, and J. H. Kim. A novel deep learning-based approach to high accuracy breast density estimation in digital mammography. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, page 101342O. International Society for Optics and Photonics, 2017.

[16] A. Akselrod-Ballin, L. Karlinsky, S. Alpert, S. Hashoul, R. Ben-Ari, and E. Barkan. A cnn based method for automatic mass detection and classification in mammograms. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–8, 2017.

[17] A. Akselrod-Ballin, L. Karlinsky, S. Alpert, S. Hasoul, R. Ben-Ari, and E. Barkan. A region based convolutional network for tumor detection and classification in breast mammography. In *Deep Learning and Data Labeling for Medical Applications*, pages 197–205. Springer, 2016.

[18] A. Akselrod-Ballin, L. Karlinsky, A. Hazan, R. Bakalo, A. B. Horesh, Y. Shoshan, and E. Barkan. Deep learning for automatic detection of abnormal findings in breast mammography. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 321–329. Springer, 2017.

[19] M. A. Al-antari, M. A. Al-masni, M.-T. Choi, S.-M. Han, and T.-S. Kim. A fully integrated computer-aided diagnosis system for digital x-ray mammograms via deep learning detection, segmentation, and classification. *International Journal of Medical Informatics*, 117:44–54, 2018.

[20] W. Al-Dhabyani, M. Gomaa, H. Khaled, and F. Aly. Deep learning approaches for data augmentation and classification of breast masses using ultrasound images. *Int. J. Adv. Comput. Sci. Appl.*, 10(5), 2019.

[21] M. Al-Masni, M. Al-Antari, J. Park, G. Gi, T. Kim, P. Rivera, E. Valarezo, S.-M. Han, and T.-S. Kim. Detection and classification of the breast abnormalities in digital mammograms via regional convolutional neural network. In *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*, pages 1230–1233. IEEE, 2017.

[22] M. A. Al-masni, M. A. Al-antari, J.-M. Park, G. Gi, T.-Y. Kim, P. Rivera, E. Valarezo, M.-T. Choi, S.-M. Han, and T.-S. Kim. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning yolo-based cad system. *Computer methods and programs in biomedicine*, 157:85–94, 2018.

[23] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *arXiv preprint arXiv:1802.06955*, 2018.

[24] N. Antropova, B. Q. Huynh, and M. L. Giger. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Medical physics*, 2017.

[25] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. G. Lopez. Convolutional neural networks for mammography mass lesion classification. In *2015 37th Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 797–800. IEEE, 2015.

[26] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. G. Lopez. Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer methods and programs in biomedicine*, 127:248–257, 2016.

[27] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

[28] I. Bakkouri and K. Afdel. Breast tumor classification based on deep convolutional neural networks. In *Advanced Technologies for Signal and Image Processing (ATSIP), 2017 International Conference on*, pages 1–6. IEEE, 2017.

[29] C. Balleyguier, S. Ayadi, K. Van Nguyen, D. Vanel, C. Dromain, and R. Sigal. Birads™ classification in mammography. *European journal of radiology*, 61(2):192–194, 2007.

[30] S. K. Bandyopadhyay. pre-processing of mammogram images. *international journal of engineering science and technology*, 2(11):6753–6758, 2010.

[31] A. S. Becker, M. Marcon, S. Ghafoor, M. C. Wurnig, T. Frauenfelder, and A. Boss. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Investigative Radiology*, 52(7):434–440, 2017.

[32] A. J. Bekker, H. Greenspan, and J. Goldberger. A multi-view deep learning architecture for classification of breast microcalcifications. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 726–730. IEEE, 2016.

[33] R. Ben-Ari, A. Akselrod-Ballin, L. Karlinsky, and S. Hashoul. Domain specific convolutional neural nets for detection of architectural distortion in mammograms. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, pages 552–556. IEEE, 2017.

[34] D. A. Berry, K. A. Cronin, S. K. Plevritis, D. G. Fryback, L. Clarke, M. Zelen, J. S. Mandelblatt, A. Y. Yakovlev, J. D. F. Habbema, and E. J. Feuer. Effect of screening and adjuvant therapy on mortality from breast cancer. *New England Journal of Medicine*, 353(17):1784–1792, 2005.

[35] R. Blanks, S. Moss, and M. Wallis. Use of two view mammography compared with one view in the detection of small invasive cancers: further results from the national health service breast screening programme. *Journal of medical screening*, 4(2):98–101, 1997.

[36] A. Bria, C. Marrocco, A. Galdran, A. Campilho, A. Marchesi, J.-J. Mordang, N. Karssemeijer, M. Molinara, and F. Tortorella. Spatial enhancement by dehazing for detection of microcalcifications with convolutional nets. In *International Conference on Image Analysis and Processing*, pages 288–298. Springer, 2017.

[37] M. L. Burman, S. H. Taplin, D. F. Herta, and J. G. Elmore. Effect of false-positive mammograms on interval breast cancer screening in a health maintenance organization. *Annals of internal medicine*, 131(1):1–6, 1999.

[38] A. Cardoso, J. Cardoso, M. J. Cardoso, I. Moreira, I. Amaral, and I. C. Domingues. Inbreast: Towards a full field digital mammographic database. 2012.

[39] G. Carneiro, D. Mateus, P. Loïc, A. Bradley, J. M. R. Tavares, V. Belagiannis, J. P. Papa, J. C. Nascimento, M. Loog, Z. Lu, et al. *Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016,*

Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings, volume 10008. Springer, 2016.

[40] G. Carneiro, J. Nascimento, and A. P. Bradley. Unregistered multiview mammogram analysis with pre-trained deep learning models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 652–660. Springer, 2015.

[41] G. Carneiro, J. Nascimento, and A. P. Bradley. Automated analysis of unregistered multi-view mammograms with deep learning. *IEEE transactions on medical imaging*, 36(11):2355–2365, 2017.

[42] G. Carneiro, J. Nascimento, and A. P. Bradley. Chapter outline. *Deep Learning for Medical Image Analysis*, page 321, 2017.

[43] N. D. G. Carneiro and A. P. Bradley. Automated mass detection from mammograms using deep learning and random forest. 2016.

[44] H.-P. Chan, S.-C. B. Lo, B. Sahiner, K. L. Lam, and M. A. Helvie. Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network. *Medical Physics*, 22(10):1555–1567, 1995.

[45] G. Chartrand, P. M. Cheng, E. Vorontsov, M. Drozdzal, S. Turcotte, C. J. Pal, S. Kadoury, and A. Tang. Deep learning: A primer for radiologists. *RadioGraphics*, 37(7):2113–2131, 2017.

[46] N. V. Chawla. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 875–886. Springer, 2009.

[47] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[48] H. Chougrad, H. Zouaki, and O. Alheyane. Convolutional neural networks for breast cancer screening: Transfer learning with exponential decay. *arXiv preprint arXiv:1711.10752*, 2017.

[49] Y. Choukroun, R. Bakalo, R. Ben-Ari, A. Akselrod-Ballin, E. Barkan, and P. Kisilev. Mammogram classification and abnormality detection from nonlocal labels using deep multiple instance neural network. 2017.

[50] I. Christoyianni, E. Constantinou, and E. Dermatas. Automatic detection of abnormal tissue in bilateral mammograms using neural networks. *Methods and Applications of Artificial Intelligence*, pages 267–275, 2004.

[51] Y.-A. Chung and W.-H. Weng. Learning deep representations of medical images using siamese cnns with application to content-based image retrieval. *arXiv preprint arXiv:1711.08490*, 2017.

[52] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013.

[53] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan. What is a good evaluation measure for semantic segmentation?. In *BMVC*, volume 27, page 2013. Citeseer, 2013.

[54] C. E. DeSantis, J. Ma, M. M. Gaudet, L. A. Newman, K. D. Miller, A. Goding Sauer, A. Jemal, and R. L. Siegel. Breast cancer statistics, 2019. *CA: a cancer journal for clinicians*, 69(6):438–451, 2019.

[55] N. Dhungel, G. Carneiro, and A. P. Bradley. Deep learning and structured prediction for the segmentation of mass in mammograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–612. Springer, 2015.

[56] N. Dhungel, G. Carneiro, and A. P. Bradley. The automated learning of deep features for breast mass classification from mammograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 106–114. Springer, 2016.

[57] N. Dhungel, G. Carneiro, and A. P. Bradley. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Medical image analysis*, 37:114–128, 2017.

[58] N. Dhungel, G. Carneiro, and A. P. Bradley. Fully automated classification of mammograms using deep residual neural networks. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, pages 310–314. IEEE, 2017.

[59] I. Domingues and J. Cardoso. Mass detection on mammogram images: a first assessment of deep learning techniques. In *19th Portuguese Conference on Pattern Recognition (RECPAD)*, 2013.

[60] I. Domingues, E. Sales, J. Cardoso, and W. Pereira. Inbreast-database masses characterization. *XXIII CBEB*, 2012.

[61] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2016.

[62] A. Dubrovina, P. Kisilev, B. Ginsburg, S. Hashoul, and R. Kimmel. Computational mammography using deep neural networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–5, 2016.

[63] N. El Atlas, M. El Aroussi, and M. Wahbi. Computer-aided breast cancer detection using mammograms: A review. In *Complex Systems (WCCS), 2014 Second World Conference on*, pages 626–631. IEEE, 2014.

[64] B. J. Erickson, P. Korfiatis, Z. Akkus, T. Kline, and K. Philbrick. Toolkits and libraries for deep learning. *Journal of Digital Imaging*, pages 1–6, 2017.

[65] M. G. Ertosun and D. L. Rubin. Probabilistic visual search for masses within mammography images using deep learning. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pages 1310–1315. IEEE, 2015.

[66] S. A. Feig. Screening mammography benefit controversies: sorting the evidence. *Radiologic Clinics of North America*, 52(3):455–480, 2014.

[67] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *International journal of cancer*, 136(5), 2015.

[68] A. Fischmann, K. Siegmann, A. Wersebe, C. Claussen, and M. Muller-Schimpfle. Comparison of full-field digital mammography and film–screen mammography: image quality and lesion detection. *The British journal of radiology*, 78(928):312–315, 2005.

[69] P. Fonseca, B. Castañeda, R. Valenzuela, and J. Wainer. Breast density classification with convolutional neural networks. In *Iberoamerican Congress on Pattern Recognition*, pages 101–108. Springer, 2016.

[70] P. Fonseca, J. Mendoza, J. Wainer, J. Ferrer, J. Pinto, J. Guerrero, and B. Castaneda. Automatic breast density classification using a convolutional neural network architecture search procedure. In *Proc. of SPIE Vol*, volume 9414, pages 941428–1, 2015.

[71] J. Gallego-Posada, D. Montoya-Zapata, and O. Quintero-Montoya. Detection and diagnosis of breast tumors using deep convolutional neural networks.

[72] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*, 2017.

[73] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70:41–65, 2018.

[74] S. J. S. Gardezi, M. Awais, I. Faye, and F. Meriaudeau. Mammogram classification using deep learning. 2017.

[75] J. Ge, L. M. Hadjiiski, B. Sahiner, J. Wei, M. A. Helvie, C. Zhou, and H.-P. Chan. Computer-aided detection system for clustered microcalcifications: comparison of performance on full-field digital mammograms and digitized screen-film mammograms. *Physics in medicine and biology*, 52(4):981, 2007.

[76] J. Ge, B. Sahiner, L. M. Hadjiiski, H.-P. Chan, J. Wei, M. A. Helvie, and C. Zhou. Computer aided detection of clusters of microcalcifications on full field digital mammograms. *Medical Physics*, 33(8):2975–2988, 2006.

[77] M. Gelig Thurfjell, B. Vitak, E. Azavedo, G. Svane, and E. Thurfjell. Effect on sensitivity and specificity of mammography screening with or without comparison of old mammograms. *Acta radiologica*, 41(1):52–56, 2000.

[78] M. J. George and S. P. Sankar. Efficient preprocessing filters and mass segmentation techniques for mammogram images. In *Circuits and Systems (ICCS), 2017 IEEE International Conference on*, pages 408–413. IEEE, 2017.

[79] K. J. Geras, S. Wolfson, S. Kim, L. Moy, and K. Cho. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv preprint arXiv:1703.07047*, 2017.

[80] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[81] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[82] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.

[83] R. C. Gonzalez, R. E. Woods, et al. Digital image processing, 2002.

[84] H. Greenspan, B. van Ginneken, and R. M. Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 2016.

[85] S. Guan and M. Loew. Breast cancer detection using synthetic mammograms from generative adversarial networks in convolutional neural networks. *Journal of Medical Imaging*, 6(3):031411, 2019.

[86] F. A. Guerrero-Pena, P. D. M. Fernandez, T. I. Ren, M. Yui, E. Rothenberg, and A. Cunha. Multiclass weighted loss for instance segmentation of cluttered cells. *arXiv preprint arXiv:1802.07465*, 2018.

[87] O. Hadad, R. Bakalo, R. Ben-Ari, S. Hashoul, and G. Amit. Classification of breast lesions using cross-modal deep learning. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, pages 109–112. IEEE, 2017.

[88] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[89] W. Hang, Z. Liu, and A. Hannun. Glimpsenet: Attentional methods for full-image mammogram diagnosis. 2017.

[90] J. H. Hayward, K. M. Ray, D. J. Wisner, J. Kornak, W. Lin, B. N. Joe, and E. A. Sickles. Improving screening mammography outcomes through comparison with multiple prior mammograms. *American Journal of Roentgenology*, 207(4):918–924, 2016.

[91] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[92] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[93] M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer. The digital database for screening mammography. *Digital mammography*, pages 431–434, 2000.

[94] M. A. Hedjazi, I. Kourbane, and Y. Genc. On identifying leaves: A comparison of cnn with classical ml methods. In *Signal Processing and Communications Applications Conference (SIU), 2017 25th*, pages 1–4. IEEE, 2017.

[95] P. U. Hepsağ, S. A. Özel, and A. Yazıcı. Using deep learning for mammography classification. In *Computer Science and Engineering (UBMK), 2017 International Conference on*, pages 418–423. IEEE, 2017.

[96] W. Ho and P. Lam. Clinical performance of computer-assisted detection (cad) system in detecting carcinoma in breasts of different densities. *Clinical Radiology*, 58(2):133–136, 2003.

[97] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau. Visual analytics in deep learning: An interrogative survey for the next frontiers. *arXiv preprint arXiv:1801.06889*, 2018.

[98] Z. Hu, J. Tang, Z. Wang, K. Zhang, L. Zhang, and Q. Sun. Deep learning for image-based cancer detection and diagnosis—a survey. *Pattern Recognition*, 2018.

[99] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[100] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger. Deep networks with stochastic depth. In *European Conference on Computer Vision*, pages 646–661. Springer, 2016.

[101] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

[102] K. Hukkinen, T. Vehmas, M. Pamilo, and L. Kivisaari. Effect of computer-aided detection on mammographic performance: experimental study on readers with different levels of experience. *Acta Radiologica*, 47(3):257–263, 2006.

[103] B. Q. Huynh, H. Li, and M. L. Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):034501–034501, 2016.

[104] S. Hwang and H.-E. Kim. Self-transfer learning for weakly supervised lesion localization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 239–246. Springer, 2016.

[105] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[106] M. M. Jadoon, Q. Zhang, I. U. Haq, S. Butt, and A. Jadoon. Three-class mammogram classification based on descriptive cnn features. *BioMed research international*, 2017, 2017.

[107] M. A. Jaffar. Deep learning based computer aided diagnosis system for breast mammograms. *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, 8(7):286–290, 2017.

[108] A. R. Jamieson, K. Drukker, and M. L. Giger. Breast image feature learning with adaptive deconvolutional networks. *SPIE Medical Imaging, Strony*, 2012:831506–831506, 2012.

[109] Y. Jia and E. Shelhamer. Caffe model zoo, 2015.

[110] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[111] F. Jiang, H. Liu, S. Yu, and Y. Xie. Breast mass lesion classification in mammograms by transfer learning. In *Proceedings of the 5th International Conference on Bioinformatics and Computational Biology*, pages 59–62. ACM, 2017.

[112] Z. Jiao, X. Gao, Y. Wang, and J. Li. A deep feature based framework for breast masses classification. *Neurocomputing*, 197:221–231, 2016.

[113] Z. Jiao, X. Gao, Y. Wang, and J. Li. A parasitic metric learning net for breast mass classification based on mammography. *Pattern Recognition*, 2017.

[114] W. Jifara, F. Jiang, S. Rho, M. Cheng, and S. Liu. Medical image denoising using convolutional neural network: a residual learning approach. *The Journal of Supercomputing*, pages 1–15, 2017.

[115] M. Kallenberg, K. Petersen, M. Nielsen, A. Y. Ng, P. Diao, C. Igel, C. M. Vachon, K. Holland, R. R. Winkel, N. Karssemeijer, et al. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE transactions on medical imaging*, 35(5):1322–1331, 2016.

[116] H. Kataoka, K. Iwata, and Y. Satoh. Feature evaluation of deep convolutional neural networks for object recognition and detection. *arXiv preprint arXiv:1509.07627*, 2015.

[117] P. Kaur and A. Kaur. Review of different approaches in mammography. 2016.

[118] D. Kinga and J. B. Adam. A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, volume 5, 2015.

[119] P. Kisilev, E. Sason, E. Barkan, and S. Hashoul. Medical image description using multi-task-loss cnn. In *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 121–129. Springer, 2016.

[120] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015.

[121] T. Kooi, B. Ginneken, N. Karssemeijer, and A. Heeten. Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network. *Medical physics*, 44(3):1017–1027, 2017.

[122] T. Kooi, A. Gubern-Merida, J.-J. Mordang, R. Mann, R. Pijnappel, K. Schuur, A. den Heeten, and N. Karssemeijer. A comparison between a deep convolutional neural network and radiologists for classifying regions of interest in mammography. In *International Workshop on Digital Mammography*, pages 51–56. Springer, 2016.

[123] T. Kooi and N. Karssemeijer. Classifying symmetrical differences and temporal change for the detection of malignant masses in mammography using deep neural networks. *Journal of Medical Imaging*, 4(4):044501, 2017.

[124] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer. Large scale deep learning for computer aided detection of mammographic lesions. *Medical image analysis*, 35:303–312, 2017.

[125] T. Kooi, J.-J. Mordang, and N. Karssemeijer. Conditional random field modelling of interactions between findings in mammography. In *SPIE Medical Imaging*, pages 101341E–101341E. International Society for Optics and Photonics, 2017.

[126] K. Krishnan, L. Baglietto, J. Stone, C. McLean, M. C. Southey, D. R. English, G. G. Giles, and J. L. Hopper. Mammographic density and risk of breast cancer by tumor characteristics: a case-control study. *BMC cancer*, 17(1):859, 2017.

[127] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[128] I. Kumar, J. Virmani, and H. Bhadauria. A review of breast density classification methods. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1960–1967. IEEE, 2015.

[129] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[130] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404, 1990.

[131] J.-G. Lee, S. Jun, Y.-W. Cho, H. Lee, G. B. Kim, J. B. Seo, and N. Kim. Deep learning in medical imaging: General overview. *Korean Journal of Radiology*, 18(4):570–584, 2017.

[132] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4:170177, 2017.

[133] C. D. Lehman, R. D. Wellman, D. S. Buist, K. Kerlikowske, A. N. Tosteson, and D. L. Miglioretti. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine*, 175(11):1828–1837, 2015.

[134] C. D. Lehman, A. Yala, T. Schuster, B. Dontchos, M. Bahl, K. Swanson, and R. Barzilay. Mammographic breast density assessment using deep learning: clinical implementation. *Radiology*, 290(1):52–58, 2018.

[135] D. Lévy and A. Jain. Breast mass classification from mammograms using deep convolutional neural networks. *arXiv preprint arXiv:1612.00542*, 2016.

[136] H. Li, M. L. Giger, B. Q. Huynh, and N. O. Antropova. Deep learning in breast cancer risk assessment: evaluation of convolutional neural networks on a clinical dataset of full-field digital mammograms. *Journal of Medical Imaging*, 4(4):041304, 2017.

[137] S. Li, J. Wei, H.-P. Chan, M. A. Helvie, M. A. Roubidoux, Y. Lu, C. Zhou, L. M. Hadjiiski, and R. K. Samala. Computer-aided assessment of breast density: comparison of supervised deep learning and feature-based statistical learning. *Physics in Medicine & Biology*, 63(2):025005, 2018.

[138] Y. Li, H. Chen, L. Cao, and J. Ma. A survey of computer-aided detection of breast cancer with mammography. *J. Health Med. Inf*, 7(4), 2016.

[139] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[140] S.-C. B. Lo, H. Li, Y. Wang, L. Kinnard, and M. T. Freedman. A multiple circular path convolution neural network system for detection of mammographic masses. *IEEE transactions on medical imaging*, 21(2):150–158, 2002.

[141] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[142] M. G. Lopez, N. Posada, D. C. Moura, R. R. Pollán, J. M. F. Valiente, C. S. Ortega, M. Solar, G. Diaz-Herrero, I. Ramos, J. Loureiro, et al. Bcdr: a breast cancer digital repository. In *15th International Conference on Experimental Mechanics*, 2012.

[143] W. Lotter, G. Sorensen, and D. Cox. A multi-scale cnn and curriculum learning strategy for mammogram classification. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 169–177. Springer, 2017.

[144] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013.

[145] A. Maier, C. Syben, T. Lasser, and C. Riess. A gentle introduction to deep learning in medical image processing. *Zeitschrift für Medizinische Physik*, 29(2):86–101, 2019.

[146] D. Masko and P. Hensman. The impact of imbalanced training data for convolutional neural networks, 2015.

[147] R. J. McKenna. The abnormal mammogram radiographic findings, diagnostic options, pathology, and stage of cancer diagnosis. *Cancer*, 74(S1):244–255, 1994.

[148] A. Melouah. Comparison of automatic seed generation methods for breast tumor detection using region growing technique. In *IFIP International Conference on Computer Science and its Applications_x000D_*, pages 119–128. Springer, 2015.

[149] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

[150] D. Mishkin, N. Sergievskiy, and J. Matas. Systematic evaluation of cnn advances on the imagenet. *arXiv preprint arXiv:1606.02228*, 2016.

[151] A. A. Mohamed, W. A. Berg, H. Peng, Y. Luo, R. C. Jankowitz, and S. Wu. A deep learning method for classifying mammographic breast density categories. *Medical Physics*, 2017.

[152] A. A. Mohamed, W. A. Berg, H. Peng, Y. Luo, R. C. Jankowitz, and S. Wu. A deep learning method for classifying mammographic breast density categories. *Medical physics*, 45(1):314–321, 2018.

[153] A. A. Mohamed, Y. Luo, H. Peng, R. C. Jankowitz, and S. Wu. Understanding clinical mammographic breast density assessment: a deep learning perspective. *Journal of Digital Imaging*, pages 1–6, 2017.

[154] J.-J. Mordang, T. Janssen, A. Bria, T. Kooi, A. Gubern-Mérida, and N. Karssemeijer. Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks. In *International Workshop on Digital Mammography*, pages 35–42. Springer, 2016.

[155] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012.

[156] A.-A. Nahid and Y. Kong. Involvement of machine learning for breast cancer image classification: a survey. *Computational and mathematical methods in medicine*, 2017, 2017.

[157] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[158] J. M. Noothout, B. D. de Vos, J. M. Wolterink, and I. Išgum. Automatic segmentation of thoracic aorta segments in low-dose chest ct. In *Medical Imaging 2018: Image Processing*, volume 10574, page 105741S. International Society for Optics and Photonics, 2018.

[159] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[160] J. E. Oliveiraa, M. O. Guelda, A. d. A. Araújob, B. Ottc, and T. M. Desernoa. Towards a standard reference database for computer-aided mammography. In *Proc. SPIE*, volume 6915, page 69151Y, 2008.

[161] A. Oliver, J. Freixenet, J. Marti, E. Perez, J. Pont, E. R. Denton, and R. Zwiggelaar. A review of automatic mass detection and segmentation in mammographic images. *Medical image analysis*, 14(2):87–110, 2010.

[162] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014.

[163] A. Oustimov, A. Gastounioti, M.-K. Hsieh, L. Pantalone, E. F. Conant, and D. Kontos. Convolutional neural network approach for enhanced capture of breast parenchymal complexity patterns associated with breast cancer risk. In *SPIE Medical Imaging*, pages 101340S–101340S. International Society for Optics and Photonics, 2017.

[164] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[165] S. Pang, Z. Yu, and M. A. Orgun. A novel end-to-end classifier using domain transferred deep convolutional neural networks for biomedical images. *Computer Methods and Programs in Biomedicine*, 140:283–293, 2017.

[166] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[167] D. Pedamonti. Comparison of non-linear activation functions for deep neural networks on mnist classification task. *arXiv preprint arXiv:1804.02763*, 2018.

[168] K. Petersen, M. Nielsen, P. Diao, N. Karssemeijer, and M. Lillholm. Breast tissue segmentation and mammographic risk scoring using deep learning. In *International Workshop on Digital Mammography*, pages 88–94. Springer, 2014.

[169] R. Platania, S. Shams, S. Yang, J. Zhang, K. Lee, and S.-J. Park. Automated breast cancer diagnosis using deep learning and region of interest detection (bc-droid). In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 536–543. ACM, 2017.

[170] D. N. Ponraj, M. E. Jenifer, P. Poongodi, and J. S. Manoharan. A survey on the preprocessing techniques of mammogram for the detection of breast cancer. *Journal of Emerging Trends in Computing and Information Sciences*, 2(12):656–664, 2011.

[171] M. Posso, M. Carles, M. Rué, T. Puig, and X. Bonfill. Cost-effectiveness of double reading versus single reading of mammograms in a breast cancer screening programme. *PloS one*, 11(7):e0159806, 2016.

[172] M. Posso, T. Puig, M. Carles, M. Rué, C. Canelo-Aybar, and X. Bonfill. Effectiveness and cost-effectiveness of double reading in digital mammography screening: A systematic review and meta-analysis. *European Journal of Radiology*, 2017.

[173] A. Qayyum, S. M. Anwar, M. Awais, and M. Majid. Medical image retrieval using deep convolutional neural network. *Neurocomputing*, 2017.

[174] Y. Qiu, Y. Wang, S. Yan, M. Tan, S. Cheng, H. Liu, and B. Zheng. An initial investigation on developing a new method to predict short-term breast cancer risk based on deep learning technology. In *SPIE Medical Imaging*, page 978521. International Society for Optics and Photonics, 2016.

[175] Y. Qiu, S. Yan, R. R. Gundreddy, Y. Wang, S. Cheng, H. Liu, and B. Zheng. A new approach to develop computer-aided diagnosis scheme of breast mass classification using deep learning technology. *Journal of X-Ray Science and Technology*, (Preprint):1–13, 2017.

[176] Y. Qiu, S. Yan, M. Tan, S. Cheng, H. Liu, and B. Zheng. Computer-aided classification of mammographic masses using the deep learning technology: a preliminary study. In *SPIE Medical Imaging*, pages 978520–978520. International Society for Optics and Photonics, 2016.

[177] V. Raman, P. Sumari, H. Then, and S. A. K. Al-Omari. Review on mammogram mass detection by machinelearning techniques. *International Journal of Computer and Electrical Engineering*, 3(6):873, 2011.

[178] R. Ramani, N. S. Vanitha, and S. Valarmathy. The pre-processing techniques for breast cancer detection in mammography images. *International Journal of Image, Graphics and Signal Processing*, 5(5):47, 2013.

[179] R. M. Rangayyan, F. J. Ayres, and J. L. Desautels. A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs. *Journal of the Franklin Institute*, 344(3-4):312–348, 2007.

[180] W. Rawat and Z. Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.

[181] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[182] R. Reiazi, R. Paydar, A. A. Ardakani, and M. Etedadialiabadi. Mammography lesion detection using faster r-cnn detector. *Computer Science & Information Technology*, page 111.

[183] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):1137–1149, 2017.

[184] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai. Detecting and classifying lesions in mammograms with deep learning. *arXiv preprint arXiv:1707.08401*, 2017.

[185] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai. Detecting and classifying lesions in mammograms with deep learning. *Scientific reports*, 8(1):4165, 2018.

[186] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[187] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[188] B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. *IEEE transactions on Medical Imaging*, 15(5):598–610, 1996.

[189] R. K. Samala, H.-P. Chan, L. Hadjiiski, M. A. Helvie, J. Wei, and K. Cha. Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. *Medical physics*, 43(12):6654–6666, 2016.

[190] R. K. Samala, H.-P. Chan, L. M. Hadjiiski, M. A. Helvie, K. H. Cha, and C. D. Richter. Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms. *Physics in Medicine & Biology*, 62(23):8894, 2017.

[191] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[192] E. Sert, S. Ertekin, and U. Halici. Ensemble of convolutional neural networks for classification of breast microcalcification from mammograms. In *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*, pages 689–692. IEEE, 2017.

[193] S. Shams, R. Platania, J. Zhang, J. Kim, K. Lee, and S.-J. Park. Deep generative breast cancer screening and diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 859–867. Springer, 2018.

[194] J. Sharma, J. Rai, and R. Tewari. Identification of pre-processing technique for enhancement of mammogram images. In *Medical Imaging, m-Health and Emerging Communication Systems (MedCom), 2014 International Conference on*, pages 115–119. IEEE, 2014.

[195] K. Sharma and B. Preet. Classification of mammogram images by using cnn classifier. In *Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on*, pages 2743–2749. IEEE, 2016.

[196] L. Shen. End-to-end training for whole image breast cancer diagnosis using an all convolutional design. *arXiv preprint arXiv:1708.09427*, 2017.

[197] P. Sherkhane and D. Vora. Survey of deep learning software tools. In *Data Management, Analytics and Innovation (ICDMAI), 2017 International Conference on*, pages 236–238. IEEE, 2017.

[198] R. L. Siegel, K. D. Miller, and A. Jemal. Cancer statistics, 2016. *CA: a cancer journal for clinicians*, 66(1):7–30, 2016.

[199] R. L. Siegel, K. D. Miller, and A. Jemal. Cancer statistics, 2017. *CA: A Cancer Journal for Clinicians*, 67(1):7–30, 2017.

[200] R. L. Siegel, K. D. Miller, and A. Jemal. Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians*, 68(1):7–30, 2018.

[201] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[202] V. K. Singh, S. Romani, H. A. Rashwan, F. Akram, N. Pandey, M. M. K. Sarker, S. Abdulwahab, J. Torrents-Barrena, A. Saleh, M. Arquez, et al. Conditional generative adversarial and convolutional networks for x-ray breast mass segmentation and shape classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 833–840. Springer, 2018.

[203] E. A. Smirnov, D. M. Timoshenko, and S. N. Andrianov. Comparison of regularization methods for imagenet classification with deep convolutional neural networks. *Aasri Procedia*, 6:89–94, 2014.

[204] A. C. Society. Breast cancer facts & figures 2017–2018, 2017.

[205] F. H. Souza, E. M. Wendland, M. I. Rosa, and C. A. Polanczyk. Is full-field digital mammography more accurate than screen-film mammography in overall population screening? a systematic review and meta-analysis. *The Breast*, 22(3):217–224, 2013.

[206] B. L. Sprague, R. E. Gangnon, V. Burt, A. Trentham-Dietz, J. M. Hampton, R. D. Wellman, K. Kerlikowske, and D. L. Miglioretti. Prevalence of mammographically dense breasts in the united states. *JNCI: Journal of the National Cancer Institute*, 106(10), 2014.

[207] S. Srinivas, R. K. Sarvadevabhatla, K. R. Mopuri, N. Prabhu, S. S. Kruthiventi, and R. V. Babu. An introduction to deep convolutional neural nets for computer vision. In *Deep Learning for Medical Image Analysis*, pages 25–52. Elsevier, 2017.

[208] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[209] W. Street. Cancer facts & figures 2019. *Am. Cancer Soc*, 76, 2018.

[210] J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S. Kok, et al. The mammographic image analysis society digital mammogram database. In *Exerpta Medica. International Congress Series*, volume 1069, pages 375–378, 1994.

[211] W. Sun, T.-L. B. Tseng, J. Zhang, and W. Qian. Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Computerized Medical Imaging and Graphics*, 57:4–9, 2017.

[212] W. Sun, T.-L. B. Tseng, B. Zheng, and W. Qian. A preliminary study on breast cancer risk analysis using deep neural network. In *International Workshop on Digital Mammography*, pages 385–391. Springer, 2016.

[213] S. Suzuki, X. Zhang, N. Homma, K. Ichiji, Y. Kawasumi, T. Ishibashi, and M. Yoshizawa. We-de-207b-02: Detection of masses on mammograms using deep convolutional neural network: A feasibility study. *Medical physics*, 43(6):3817–3817, 2016.

[214] S. Suzuki, X. Zhang, N. Homma, K. Ichiji, N. Sugita, Y. Kawasumi, T. Ishibashi, and M. Yoshizawa. Mass detection using deep convolutional neural network for mammographic computer-aided diagnosis. In *Society of Instrument and Control Engineers of Japan (SICE), 2016 55th Annual Conference of the*, pages 1382–1386. IEEE, 2016.

[215] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[216] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[217] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. On the necessity of fine-tuned convolutional neural networks for medical imaging. In *Deep Learning and Convolutional Neural Networks for Medical Image Computing*, pages 181–193. Springer, 2017.

[218] P. Teare, M. Fishman, O. Benzaquen, E. Toledano, and E. Elnekave. Malignancy detection on mammography using dual deep convolutional neural networks and genetically discovered false color input enhancement. *Journal of Digital Imaging*, 30(4):499–505, 2017.

[219] J. Teuwen, S. van de Leemput, A. Gubern-Mérida, A. Rodriguez-Ruiz, R. Mann, and B. E. Bejnordi. Soft tissue lesion detection in mammography using deep neural networks for object detection. 2018.

[220] R. L. Thomaz, P. C. Carneiro, and A. C. Patrocinio. Feature extraction using convolutional neural network for classifying breast density in mammographic images. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, page 101342M. International Society for Optics and Photonics, 2017.

[221] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[222] K. Umehara, J. Ota, and T. Ishida. Super-resolution imaging of mammograms based on the super-resolution convolutional neural network. *Open Journal of Medical Imaging*, 7(04):180, 2017.

[223] G. Valvano, D. Della Latta, N. Martini, G. Santini, A. Gori, C. Iacconi, A. Ripoli, L. Landini, and D. Chiappino. Evaluation of a deep convolutional neural network method for the segmentation of breast microcalcifications in mammography imaging. In *EMBEC & NBC 2017*, pages 438–441. Springer, 2017.

[224] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[225] A. Vedaldi and K. Lenc. Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 689–692. ACM, 2015.

[226] D. M. Vo and S.-W. Lee. Semantic image segmentation using fully convolutional neural networks with multi-scale images and multi-scale dilated convolutions. *Multimedia Tools and Applications*, pages 1–19, 2018.

[227] A. T. Wang, C. M. Vachon, K. R. Brandt, and K. Ghosh. Breast density and breast cancer risk: a practical review. In *Mayo Clinic Proceedings*, volume 89, pages 548–557. Elsevier, 2014.

[228] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.

[229] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.

[230] J. Wang, H. Ding, F. Azamian, B. Zhou, C. Iribarren, S. Molloi, and P. Baldi. Detecting cardiovascular disease from mammograms with deep learning. *IEEE transactions on medical imaging*, 2017.

[231] K. Wang, Y. He, H. Qin, P. R. Fisher, and W. Zhao. Temporal registration of 2d x-ray mammogram using triangular b-splines finite element method (tbfem). In *Medical Imaging 2006: Image Processing*, volume 6144, page 614436. International Society for Optics and Photonics, 2006.

[232] D. Wei, B. Sahiner, H.-P. Chan, and N. Petrick. Detection of masses on mammograms using a convolution neural network. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 5, pages 3483–3486. IEEE, 1995.

[233] X. Wei, J. Chen, and C. Cai. Using deep convolutional neural networks and transfer learning for mammography mass lesion classification. *Journal of Computational and Theoretical Nanoscience*, 14(8):3802–3806, 2017.

[234] K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, 2016.

[235] H. G. Welch and H. J. Passow. Quantifying the benefits and harms of screening mammography. *JAMA internal medicine*, 174(3):448–454, 2014.

[236] I. Wichakam and P. Vateekul. Combining deep convolutional networks and svms for mass detection on digital mammograms. In *Knowledge and Smart Technology (KST), 2016 8th International Conference on*, pages 239–244. IEEE, 2016.

[237] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum. Dilated convolutional neural networks for cardiovascular mr segmentation in congenital heart disease. In *Reconstruction, Segmentation, and Analysis of Medical Images*, pages 95–102. Springer, 2016.

[238] N. Wu, K. J. Geras, Y. Shen, J. Su, S. Kim, E. Kim, S. Wolfson, L. Moy, and K. Cho. Breast density classification with deep convolutional neural networks. *arXiv preprint arXiv:1711.03674*, 2017.

[239] N. Wu, K. J. Geras, Y. Shen, J. Su, S. G. Kim, E. Kim, S. Wolfson, L. Moy, and K. Cho. Breast density classification with deep convolutional neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6682–6686. IEEE, 2018.

[240] P. Xi, C. Shu, and R. Goubran. Abnormality detection in mammography using deep convolutional neural networks. *arXiv preprint arXiv:1803.01906*, 2018.

[241] J. Xu, C. Li, Y. Zhou, L. Mou, H. Zheng, and S. Wang. Classifying mammographic breast density by residual learning. *arXiv preprint arXiv:1809.10241*, 2018.

[242] S. Yadav and S. Shukla. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In *Advanced Computing (IACC), 2016 IEEE 6th International Conference on*, pages 78–83. IEEE, 2016.

[243] N. I. Yassin, S. Omran, E. M. El Houby, and H. Allam. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer Methods and Programs in Biomedicine*, 2017.

[244] D. Yi, R. L. Sawyer, D. Cohn III, J. Dunnmon, C. Lam, X. Xiao, and D. Rubin. Optimizing and visualizing deep learning for benign/malignant classification in breast tumors. *arXiv preprint arXiv:1705.06362*, 2017.

[245] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[246] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[247] F. Yu, V. Koltun, and T. A. Funkhouser. Dilated residual networks. In *CVPR*, volume 2, page 3, 2017.

[248] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[249] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[250] L. Zhang, L. Lin, X. Liang, and K. He. Is faster r-cnn doing well for pedestrian detection? In *European conference on computer vision*, pages 443–457. Springer, 2016.

[251] X. Zhang, Y. Zhang, E. Y. Han, N. Jacobs, Q. Han, X. Wang, and J. Liu. Whole mammogram image classification with convolutional neural networks. In *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*, pages 700–704. IEEE, 2017.

[252] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao. Deep learning and its applications to machine health monitoring: A survey. *arXiv preprint arXiv:1612.07640*, 2016.

[253] Y. Zheng, C. Yang, and A. Merkulov. Breast cancer screening using convolutional neural network and follow-up digital mammography. In *Computational Imaging III*, volume 10669, page 1066905. International Society for Optics and Photonics, 2018.

[254] Y. Zheng, C. Yang, A. Merkulov, and M. Bandari. Early breast cancer detection with digital mammograms using haar-like features and adaboost algorithm. In *Sensing and Analysis Technologies for Biomedical and Cognitive Applications 2016*, volume 9871, page 98710D. International Society for Optics and Photonics, 2016.

[255] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018.

[256] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 603–611. Springer, 2017.

[257] W. Zhu, X. Xiang, T. D. Tran, G. D. Hager, and X. Xie. Adversarial deep structured nets for mass segmentation from mammograms. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 847–850. IEEE, 2018.

[258] W. Zhu and X. Xie. Adversarial deep structural networks for mammographic mass segmentation. *arXiv preprint arXiv:1612.05970*, 2016.

[259] L. Zou, S. Yu, T. Meng, Z. Zhang, X. Liang, and Y. Xie. A technical review of convolutional neural network-based mammographic breast cancer diagnosis. *Computational and Mathematical Methods in Medicine*, 2019, 2019.

[260] K. Zuiderveld. Contrast limited adaptive histogram equalization. *Graphics gems*, pages 474–485, 1994.